IRE Transactions on INFORMATION THEORY

A Journal Devoted to the Theoretical and Experimental Aspects of Information Transmission, Processing and Utilization.

Volume IT-7 JANUARY, 1961 Number 1
Published Quarterly

In This Issue

PUBLISHED BY THE
Professional Group on Information Theory

IRE Professional Group on Information Theory

The Professional Group on Information Theory is an organization, within the framework of the IRE, of members with principal professional interest in Information Theory. All members of the IRE are eligible for membership in the Group and will receive all Group publications upon payment of an annual fee of $4.00.

IRE TRANSACTIONS® on INFORMATION THEORY is published in January, April, July, and October, by the IRE for the Professional Group on Information Theory, at 1 East 79th Street, New York 21, N. Y. In addition to these regular quarterly issues, Special Issues appear from time to time. Responsibility for contests rest upon the authors and not upon the IRE, the Group, or its members. Individual copies of this issue and all available back issues, except PGIT-4, may be purchased at the following prices: IRE members (one copy) $2.25, libraries and colleges $3.25, all others $4.50.

INFORMATION THEORY

IRE Transactions on Information Theory

A Journal Devoted to the Theoretical and Experimental Aspects of Information Transmission, Processing and Utilization

Volume IT-7 | January, 1961 | Number 1
Published Quarterly

TABLE OF CONTENTS

Contributions

Correspondence

Single Error-Correcting Codes for Nonbinary Balanced Channels*

C. W. HELSTROM†

Summary—**Close-packed, single error-correcting codes are studied, the letters of which are N-tuples of M-ary digits, where M is the power of a prime. The length N of the letters must be given by $N = (M^k - 1)/(M - 1)$, where k is an integer. A balanced communication channel, for which all errors in a transmitted digit are equally likely, is defined and a physical model given. The probability of correct reception of the code letters and the rate with which they transmit information in a balanced channel are calculated. This involves deriving formulas for the numbers of code letters having various numbers of 0's. Numerical results are given for a quaternary code and are extended to the case where the quaternary channel has a null zone, so that erasures as well as errors may occur. For the type of signals and noise assumed, the balanced channel without a null zone is found to yield the better performance.**

I. Description of the Codes

THE LETTERS of the code alphabets to be considered here are sets of N digits, each of which can take on any of M values. For the most part we shall assume that M is an integral power of a prime number p and that $M > 2$. We shall restrict ourselves to close-packed codes that correct all single errors in sets of N received digits, an error being defined as any alteration of a transmitted digit. The length of each code letter is given by

$$N = \frac{M^k - 1}{M - 1}, \tag{1}$$

where k is any integer greater than 1, and the alphabet contains M^{N-k} letters. The term "close-packed" means that any of the M^N possible sets of N digits either is a code letter or differs from a code letter in only one place. We shall describe some of the properties of these codes and show how to evaluate their performance in a particularly simple type of communication channel.

The existence of these codes for M (the power of a prime) was demonstrated by Zaremba,[1] who used the methods of group theory. Such codes for M prime have been studied by Ulrich[2] and Shapiro and Slotnick.[3] An example of this kind of code for $M = 4$, $k = 2$ was given by Golay.[4]

Cocke[5] has shown how close-packed, single error-correcting codes can be constructed by associating the several digits with the elements of a Galois field $GF(M)$ of order M, where M is any power of a prime number, $M = p^r$. Under the operation of addition, the elements of the field form an Abelian group of type[6] $(p, p, \cdots, p)$ (r times). That is, there are r elements such that the $M = p^r$ field elements can be generated by adding them to each other in all possible ways. Each of these r basis elements is of order p under addition. Here "1" is the identity element under field multiplication. The multiplicative group is also Abelian, but cyclic and of order $p^r - 1$. The code letters are N-tuples of elements of the Galois field. One code letter will be the N-tuple $I = (0, 0, \cdots, 0)$ consisting of all 0's, where 0 is the identity element under the field operation of addition. All the rest of the code letters must have at least three nonzero digits in order for single errors to be uniquely correctable.[3]

The digits y_j of each code letter satisfy k linear relations of the form

$$\sum_{j=1}^{N} a_{ij} y_j = 0, \qquad i = 1, 2, \cdots, k, \tag{2}$$

where the a_{ij} are field elements, and the additions and multiplications are carried out by the rules for the Galois field. These equations are analogous to the parity-check relations for binary codes. Cocke[5] has shown that if k is linearly independent such relations can always be found, and that the resulting code will correct all single errors.

As an example, we refer to the code listed by Golay[4] in (10) of his paper. The field of order 4 can be described by the relations[5]

$$x + x = 0, \qquad \alpha \cdot \alpha^2 = 1, \qquad 1 + \alpha + \alpha^2 = 0, \qquad \alpha^3 = 1,$$

where x is any of the four elements 0, 1, α, and α^2. In Golay's code letters we replace the digits "2" and "3" by α and α^2 respectively. Then the digits of those letters satisfy the linear relations

$$\begin{aligned} x_1 + x_2 + x_3 + x_4 &= 0, \\ \alpha x_1 + x_2 + \alpha^2 x_3 + x_5 &= 0. \end{aligned} \tag{3}$$

* Received by the PGIT, December 1, 1959; revised manuscript received, August 17, 1960. Westinghouse Res. Labs., Pittsburgh, Pa., Scientific Paper 412FF471-P1.

† Dept. of Math., Westinghouse Res. Labs., Pittsburgh, Pa.

[1] S. K. Zaremba, "Covering problems concerning Abelian groups," *J. London Math. Soc.*, vol. 27, pp. 242–246; April, 1952.

[2] W. Ulrich, "Non-binary error correction codes," *Bell Sys. Tech. J.*, vol. 36, pp. 1341–1388; November, 1957.

[3] H. S. Shapiro and D. L. Slotnick, "On the mathematical theory of error-correcting codes," *IBM J. Res. Dev.*, vol. 3, pp. 25–34; January, 1959.

[4] M. J. E. Golay, "Notes on the penny-weighing problem, lossless symbol coding with nonprimes, etc." IRE Trans. on Information Theory, vol. IT-4, pp. 103–109; September, 1958.

[5] J. Cocke, "Lossless symbol coding with nonprimes," IRE Trans. on Information Theory, vol. IT-5, pp. 33–34; March, 1959.

[6] A. Speiser, "Die Theorie der Gruppen von endlicher Ordnung," Dover Publications, Inc., New York, N. Y.; 1945. See ch. 3, pp. 46–64.

quaternary code equivalent to this code was independently discovered by Scherer.[7] We shall evaluate its performance in a later section of this paper.

The letters of these codes are themselves elements of an Abelian group of order M^{N-k}, which in turn is a subgroup of the group of order M^N containing all possible N-tuples of M-ary digits. The group operation is digit-by-digit (vectorial) field addition. Following Slepian[8] we can divide the latter group into M^k cosets by "factoring" out the subgroup of code letters. We take as the "leader" of each coset the N-tuple with the greatest number of 0's; these coset leaders are just the $(M-1)N$ N-tuples with $(N-1)$ 0's and with one of the $(M-1)$ nonzero field elements in the remaining place, along with the identity element $I = (0, 0, \cdots 0)$ of the code group. No two of these N-tuples can occur in the same coset, for then their difference would be a code letter different from I and having fewer than three nonzero elements, an impossibility if all single errors are to be uniquely correctable.) We can then write out the code letters in a horizontal line, and under each letter put its sums with the $(M-1)N$ coset leaders other than I. Since for N given by (1) this procedure exhausts the set of M^N possible received N-tuples, the code is close-packed. Each received set is found in one and only one column, and it is either the head of that column or differs from it in only one place. Each such N-tuple is to be decoded into the letter at the head of the column in which it is found.

Given a received set $(u_1, u_2, \cdots u_N)$, one can form a k-tuple $(\gamma_1, \gamma_2, \cdots \gamma_k)$, where

$$\gamma_i = \sum_{j=1}^{N} a_{ij} u_j. \tag{4}$$

is known as a "corrector".[2] All elements of a coset have the same corrector. A received N-tuple can be decoded by calculating its corrector, looking up in a table the associated coset leader, and adding the inverse (or "negative") of the coset leader to the N-tuple. When M is prime,[2] the matrix a_{ij} can be so chosen that the corrector specifies immediately which digit is to be changed and by how much.

II. Calculation of Code Performance in a Balanced Channel

A balanced channel is defined in terms of its conditional probabilities $p_i(j)$ of receiving the digit j when the digit i is transmitted. The probability $\beta = p_i(i)$ of correct reception is the same for all digits, and the probabilities $\delta = p_i(j)$ are equal for all $j \neq i$ and are independent of i. Errors in successive digits are assumed to be statistically independent. A physical model of such a channel will be described later.

In the balanced channel the probability $p_X(Y)$ of receiving the N-tuple Y when the code letter X was transmitted has the following useful symmetry property:

$$p_X(Y) = p_I(Y - X) \tag{5}$$

where I is the N-tuple of all 0's, and $Y - X = Y + (-X)$ where $(-X)$ is the element inverse to X under the operation of group addition $X + (-X) = I$. All code letters are transmitted with equal relative frequencies. For such a channel the decoding procedure described above always picks the code letter with largest posterior probability. To evaluate the performance of our codes in this kind of channel we can use Slepian's rules[8] for computing the probability $\bar{Q}$ of correct reception of a transmitted letter and the rate R of transmission of information. Although Slepian derived them for the binary symmetric channel, they depend only on the symmetry (5) and on the group properties of the code, and they are therefore applicable here.

To apply these rules one takes the table of M^N received N-tuples, described above, and assigns to each member of it the probability $\beta^m \delta^{N-m}$ that it is received when the letter $I = (0, 0, \cdots 0)$ is sent, where m is the number of 0's in the N-tuple. One then forms the column sum Q_X by adding these probabilities for all N-tuples in the column headed by the code letter X, including that for X itself. The probability $\bar{Q}$ of correct reception is given by the sum Q_I for the first column, and it is easily seen to be

$$\bar{Q} = Q_I = \beta^N + (M-1)N\beta^{N-1}\delta. \tag{6}$$

The rate R of transmission is

$$R = \frac{1}{N}\left[(N-k)\log M + \sum_X Q_X \log Q_X\right] \tag{7}$$

bits per digit, where we use logarithms to the base 2. In (7) the sum is taken over all the code letters X.

This sum is not as formidable as it appears, for many of the Q_X's are equal. Indeed, we have

$$Q_X = \pi_{m(X)} \tag{8}$$

where $m = m(X)$ is the number of 0's in the code letter X. If we let ν_m be the number of code letters with m 0's, the rate of transmission is

$$R = \frac{1}{N}\left[(N-k)\log M + \sum_{m=0}^{N} \nu_m \pi_m \log \pi_m\right]. \tag{9}$$

We have $\nu_N = 1$, $\pi_N = Q_I$ of (6).

To compute Q_X we must first count up, in the column under the letter X, the number of N-tuples having various numbers of 0's. We recall that they are formed by adding the coset leaders to the letter X. We can form sets having one fewer 0 than X by adding any of the $(M-1)$ nonzero field elements to any of the m places of X occupied by 0. There are $m(M-1)$ ways this can be done. The sets having one more zero than X are formed by adding

[7] R. Filipowsky, P. Portmann, and E. Scherer, "Improvements obtained by a Quaternary Code and Decision System," presented at the Third AeroCom Sym., Rome-Utica, N. Y.; November 8, 1957.

[8] D. Slepian, "A class of binary signaling alphabets," *Bell Sys. Tech. J.*, vol. 35, pp. 203–234; January, 1956.

to each nonzero element x of X its inverse $(-x)$ under addition. This can be done in $(N - m)$ ways. Sets with the same number of 0's as X can be formed by adding to the $(N - m)$ nonzero elements x of X any of the $(M - 2)$ field elements differing from both $(-x)$ and 0. Including X there are $[(N - m)(M - 2) + 1]$ sets with m 0's in the column. Thus we get the column sum

$$\begin{aligned} Q_X = \pi_{m(X)} &= m(M-1)\beta^{m-1}\,\delta^{N-m+1} \\ &\quad + [(M-2)(N-m)+1]\beta^{m}\,\delta^{N-m} \\ &\quad + (N-m)\beta^{m+1}\,\delta^{N-m-1}, \\ m &= m(X). \end{aligned} \tag{10}$$

This way of calculating the column sum Q_X shows that it depends only on the number m of 0's in X, and neither on the remaining digits of X nor on their arrangement.

III. The Numbers ν_m.

It remains only to calculate the number ν_m of code letters with m 0's. The ν_m's are important not only for calculating the rate of transmission, but also for describing the structure of the group of code letters. They have been computed by Lloyd[9] for close-packed binary codes, and it is a simple matter to apply his method to our M-ary codes.

We define ν_m^1 to be the number of elements having m 0's and differing from some code letter in one digit. The total number of N-tuples with m 0's is

$$\nu_m + \nu_m^1 = \binom{N}{m}(M-1)^{N-m}. \tag{11}$$

The sets with m 0's and numbered by ν_m^1 are formed from code letters with $(m - 1)$, m, and $(m + 1)$ 0's by adding field elements in the proper places, much as we generated the elements in a column in calculating Q_X. It is easy to modify that development to show that

$$\begin{aligned} \nu_m^1 &= (M-1)(m+1)\nu_{m+1} \\ &\quad + (N-m)(M-2)\nu_m + (N-m+1)\nu_{m-1}. \end{aligned} \tag{12}$$

Combining (11) and (12) we get the set of difference equations

$$\begin{aligned} &(M-1)(m+1)\nu_{m+1} + [(N-m)(M-2)+1]\nu_m \\ &\quad + (N-m+1)\nu_{m-1} = \binom{N}{m}(M-1)^{N-m}. \end{aligned} \tag{13}$$

They can be most easily solved by introducing the generating function[9]

$$G(z) = \sum_{s=0}^{N} \nu_s z^s. \tag{14}$$

[9] S. Lloyd, "Binary block coding," *Bell Sys. Tech. J.*, vol. 36, pp. 517–535; March, 1957.

Using the difference equations (13) one can then derive the following differential equation for $G(z)$:

$$\begin{aligned} &[M - 1 - (M-2)z - z^2]\frac{dG}{dz} \\ &\quad + [N(M-2) + 1 + Nz]G(z) = (M - 1 + z)^N. \end{aligned} \tag{15}$$

Since we know that $\nu_N = 1$, we look for a solution that behaves like z^N as z approaches infinity. Independently of any relation between M and N, this solution is

$$\begin{aligned} G(z) &= \frac{(z + M - 1)^B}{N(M-1)+1} \\ &\quad \cdot[(z + M - 1)^A + N(M-1)(z-1)^A], \\ A &= [N(M-1)+1]/M, \qquad B = (N-1)/M, \\ &\qquad A + B = N. \end{aligned} \tag{16}$$

This result agrees with Lloyd's[9] for $M = 2$, $N = 2^k - 1$.

The total number of code letters is

$$G(1) = \sum_{s=0}^{N} \nu_s = \frac{M^N}{N(M-1)+1}, \tag{17}$$

which must be a power of M. Hence the length N of the code letters must be given by a formula like (1). If we expand $G(z)$ in powers of z we get

$$G(z) = z^N + \tfrac{1}{6}N(N-1)(M-1)^2 z^{N-3} + 0(z^{N-4}). \tag{18}$$

Therefore $\nu_{N-1} = \nu_{N-2} = 0$, and all code letters except I have at least three nonzero digits. The general term in the expansion of $G(z)$ yields the number of code letters with m zeros:

$$\begin{aligned} \nu_m &= \frac{1}{AM}\left\{\binom{N}{m}(M-1)^{N-m} + N(M-1)\right. \\ &\quad \left.\cdot\sum_{r=0}^{B}\binom{B}{r}\binom{A}{m-r}(M-1)^{B-r}(-1)^{A-r+m}\right\}, \end{aligned} \tag{19}$$

provided one takes the binomial coefficient to equal 0 when its lower index is negative. In the simplest case of $k = 2$, $N = M + 1$, we have $B = 1$, $A = M$, and

$$\begin{aligned} \nu_m &= M^{-2}\binom{N}{m}[(M-1)^{N-m} \\ &\quad - (-1)^{N-m}(M-1)(M^2 - 1 - Mm)]. \end{aligned} \tag{20}$$

The results of this section do not depend on M being a power of a prime, and one is led to speculate whether close-packed, single-error correcting codes exist for composite integers M. That a *group* code of this kind cannot exist for $M = 6$, $k = 2$, $N = 7$ becomes apparent when one tries to set up a basis for it. An Abelian group of order 6 has one and only one element of order 2. If we call it a, $a + a = 0$. Now consider the digit sets $(a, 0, 0, 0, 0, x, y)$ and $(0, a, 0, 0, 0, z, w)$. Since for these to be code letters they must have three nonzero elements, none of x, y, z, and w can be 0. Each of these two septuples when added to itself must yield the identity letter I, for otherwise there would be code letters other than I with

ver than three nonzero elements. Therefore x, y, z, d w must be of order 2, and $x = y = z = w = a$. If then add the two septuples together, we get a code ter with two nonzero elements, an impossibility.

If for a general composite integer M we consider ele-nts of the code group of the form

$$\underbrace{(a, 0, 0, \cdots, 0}_{N-k}, \underbrace{x, y, z, \cdots)}_{k},$$

d let a run through the elements of an Abelian group order M, the digits in each place of the k-tuple on the ht run through an automorphism of the same Abelian oup. One may be able to draw conclusions about the stence of close-packed, single-error correcting group les for M-ary channels, M composite, by studying the tomorphisms of the Abelian groups of order M, keeping mind that at least two of the elements of each k-tuple the right must differ from 0, except in the identity ter I.

IV. A Model of a Balanced Channel

A transmitter can send any of M narrowband signals duration T and of the form

$$\text{real part of } F_j(t)e^{iWt}, \qquad 1 \leq j \leq M,$$

ere W is 2π times the carrier frequency. The complex velopes $F_j(t)$ of the signals are orthogonal, as follows:

$$\int_0^T F_j^*(t)F_k(t)\,dt = E_0\,\delta_{jk}; \tag{21}$$

s is merely a way of specifying that the signals do t "overlap." All signals are received with equal energies, t corrupted by white Gaussian noise and with unknown rrier phases. The signals are passed through a set of parallel filters, each matched to one of the signals in e sense of detection theory,[10] and each followed by a ear detector. The outputs of the M detectors are easured at the end of each transmission interval of ration T. Suitably normalized, these outputs are de-ted by $(y_1, y_2, \cdots y_M)$ and are described by the prob-lity density functions[10]

$$) = q(0, y) = ye^{-y^2/2} \quad \text{(signal } j \text{ absent)}$$

$$) = q(d, y) = ye^{-(y^2+d^2)/2}I_0(dy) \quad \text{(signal } j \text{ present)} \tag{22}$$

$$y = y_i > 0, \qquad 1 \leq j \leq M.$$

e outputs are statistically independent. Here $d = \sqrt{2E/N_0}$ is defined as the signal-to-noise ratio, where E he energy of the received signal pulses and N_0 is the ctral density of the noise.

The M outputs $(y_1, \cdots y_M)$ are fed to a "decider," which its the digit corresponding to the largest of them. For

[10] C. W. Helstrom, "Statistical Theory of Signal Detection," gamon Press, Ltd., London, Eng.; 1960. See ch. 5, pp. 129–165.

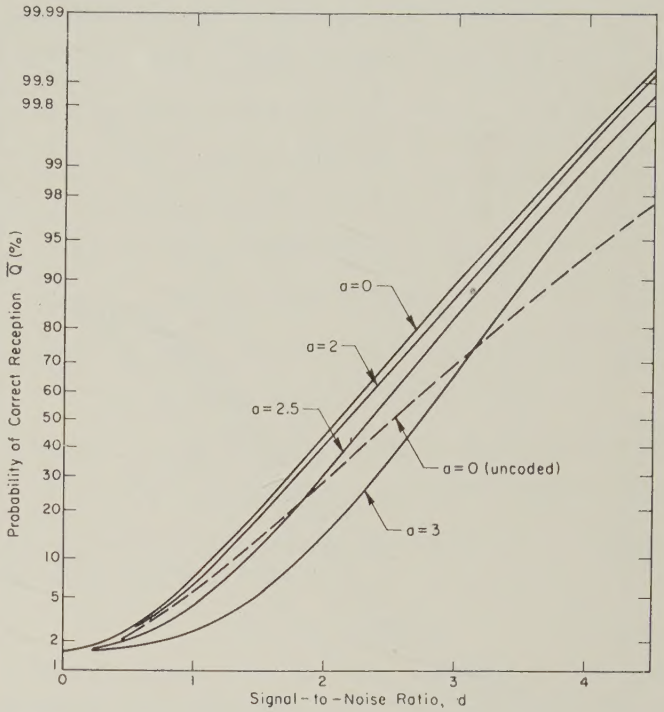


Fig. 1—Probability of correct reception for quaternary code.

this channel the conditional probabilities are[11]

$$\beta = p_i(i) = \int_0^\infty q(d, y_1)\,dy_1$$

$$\cdot \int_0^{y_1} q(0, y_2)\,dy_2 \cdots \int_0^{y_1} q(0, y_M)\,dy_M$$

$$= \sum_{r=0}^{M-1} \frac{(-1)^r}{r+1}\binom{M-1}{r} \exp\,[-rd^2/2(r+1)],$$

$$\delta = p_i(j) = (1-\beta)/(M-1), \qquad j \neq i.$$

We have calculated the probability $\bar{Q}$ of correct reception and the rate R of transmission for a quaternary code[4,7] for which $M = 4$, $k = 2$, $N = 5$. For this code,

$$\nu_5 = 1, \quad \nu_4 = \nu_3 = 0, \quad \nu_2 = 30, \quad \nu_1 = 15, \quad \nu_0 = 18. \tag{24}$$

We used the above-described model for the balanced channel, with $M = 4$. The results are plotted in Figs. 1 and 2 as the solid curves marked $a = 0$. For purposes of comparison we have plotted as dashed lines the corresponding results for a system transmitting the 64 sets of three quaternary digits without coding. For these the probability of correct reception is simply $\bar{Q} = \beta^3$, and the rate of transmission is equal to the capacity of the balanced channel, as follows:

$$R = \log M + \beta \log \beta + (M-1)\,\delta \log \delta \tag{25}$$

[11] C. W. Helstrom, "The Performance of Communication Channels with Orthogonal Signals," Westinghouse Elec. Corp. Res. Rept. 412FF471–R1; August 28, 1959. See also S. Reiger, "Error rates in data transmission," Proc. IRE, vol. 46, p. 919; May, 1958.

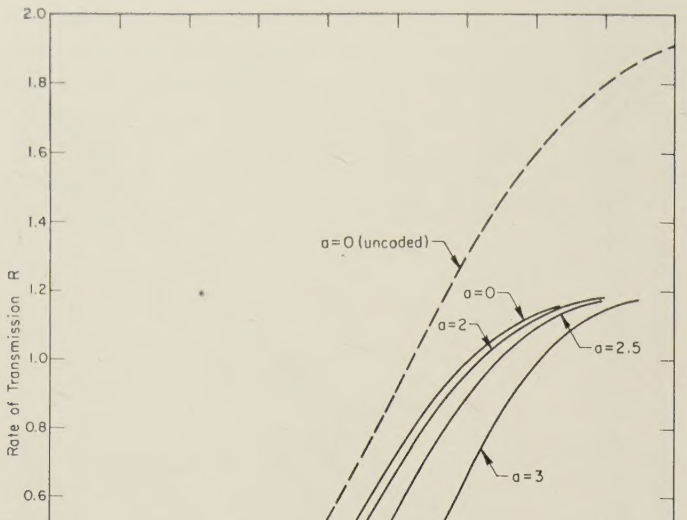


Fig. 2—Rate of transmission for quaternary code.

bits per digit. The coded digit sets have a higher probability of being correctly received, but they convey information at a smaller rate than the uncoded sets.

V. The Balanced Channel with Null Zone

The use of null zones to increase channel capacity has been discussed by Bloom, *et al.*,[12] for the binary symmetric channel. A simple kind of null zone for the M-ary balanced channel can be set up by instructing the decider to emit an erasure, denoted by "E," whenever all M detector outputs $y_1, \cdots y_M$ fall below a certain threshold $y = a$. Otherwise the decider emits the digit corresponding to the largest output. The transition probabilities for this new channel are as follows:[11]

$$\beta = p_i(i) = \int_a^\infty q(d, y)\, dy \left(\int_0^y q(0, x)\, dx \right)^{M-1}$$

$$= \sum_{r=0}^{M-1} \frac{(-1)^r}{r+1} \exp\left[-r\, d^2/2(r+1)\right]$$

$$\cdot Q\left(\frac{d}{\sqrt{r+1}}, a\sqrt{r+1} \right),$$

$$\gamma = p_i(E) = \int_0^a q(d, y)\, dy \left(\int_0^a q(0, x)\, dx \right)^{M-1}$$

$$= [1 - Q(d, a)](1 - e^{-a^2/2})^{M-1}$$

$$\delta = p_i(j) = (1 - \beta - \gamma)/(M - 1), \quad j \neq i, \; j \neq E. \tag{26}$$

Here d is again the signal-to-noise ratio, and

$$Q(\alpha, \beta) = \int_\beta^\infty x e^{-(x^2+\alpha^2)/2} I_0(\alpha x)\, dx \tag{27}$$

is Marcum's Q-function.[13]

There exists a value $\bar{a}$ of the threshold for which the channel capacity

$$C(a) = (1 - \gamma) \log \left(\frac{M}{1 - \gamma} \right) + \beta \log \beta + (M - 1)\, \delta \log \delta \tag{28}$$

is a maximum. This maximum capacity $C(\bar{a})$ was found to be only slightly larger than $C(0)$, the capacity of a channel without a null zone, for the cases $M = 2$, 4 studied.[11]

Since the single-error correcting codes described above also correct double erasures, it might be thought that they would yield a better performance with some non-vanishing null zone than with $a = 0$. To test this, we have calculated the probability $\bar{Q}$ of correct reception and the rate R of transmission when the quaternary code ($M = 4$, $N = 5$, $k = 2$) is applied to such a balanced channel with null zone.[14] Using a digital computer, these were evaluated for a range of values of d and for $a = 0$ to 3 in steps of 0.5. The results are plotted in Figs. 1 and 2. (The curves for $a = 0.5$, 1.0, 1.5 lie between those for $a = 0$ and $a = 2$.) It was found that both $\bar{Q}$ and R decrease as the threshold a increases, slowly at first, and then more rapidly as a exceeds d. Thus for $M = 4$, with the type of signals and noise considered here, it is preferable not to use a null zone when applying this kind of code, even though it does correct double erasures.

In our analysis[14] we postulated a decoding procedure that chooses the code letter with the greatest posterior probability in view of the digits emitted by the decider. For received sets with no erasures it is the same as that described above. For sets with two erasures the missing digits are calculated from the linear relations (4) for the code ($k = 2$).

If a set contains a single erasure, the remaining four digits may or may not be part of some code letter. If they are, that letter has maximum posterior probability and is emitted by the decoder. If they are not, there are four code letters with equal and largest posterior probabilities. We assumed that the decoder picks one of these four with probability 1/4. Thus a set with one erasure can be decoded by filling in the erasure with a quaternary digit picked at random, after which the decoding procedure for sets with no erasures is applied. If three or more erasures occur, all but two are filled in with quaternary digits chosen independently and at random, and the remaining two digits are calculated from the linear check relations (4).

[12] F. J. Bloom, *et al.*, "Improvement of binary transmission by null-zone reception," Proc. IRE, vol. 45, pp. 963–975; July, 1957.

[13] J. I. Marcum, "A Table of Q-Functions," Rand Corp., Rept. RM-339; January 1, 1950.

Under this decoding procedure the received sets retain he symmetry property (5) required for the validity of lepian's rules for calculating $\bar{Q}$ and R. It is then a matter f listing all possible received N-tuples, including those ith one to five erasures, underneath the code letters nto which they are decoded. To each is assigned the robability $\beta^m\gamma^n\delta^{N-m-n}$ of being received when the letter $= (0, 0, \cdots 0)$ is sent, where m is the number of 0's nd n the number of E's in the N-tuple. The column ums Q_X are then formed as before, but with one modifiation. For those N-tuples involving erasures and decoded y a chance device, the probability $\beta^m\gamma^n\delta^{N-m-n}$ is weighted ith the probability that the letter X is picked by the ecoder. These weights are 1/4 for $n = 1$ and 3, 1/16 or $n = 4$, and 1/64 for $n = 5$. Since the sum Q_X again epends only on the number of 0's in the code letter X, ne can use (9) for the rate of transmission.

The tedious task of counting up all the terms is described elsewhere.[14] We only list the column sums for he case $M = 4$, $N = 5$:

$$\begin{aligned}\pi_5 &= \beta^5 + 15\beta^4\delta + 5\beta^4\gamma + 60\beta^3\gamma\delta/4 \\ &\quad + 10\beta^3\gamma^2 + 10\beta^2\gamma^3/4 + 5\beta\gamma^4/16 + \gamma^5/64.\end{aligned}$$

$$\begin{aligned}\pi_2 &= 3\beta^3\delta^2 + 7\beta^2\delta^3 + 6\beta\delta^4 + \gamma(2\beta\delta^3 + 3\beta^2\delta^2) \\ &\quad + (6\beta^3\delta + 18\beta^2\delta^2 + 30\beta\delta^3 + 6\delta^4)\gamma/4 \\ &\quad + (3\beta^2\delta + 6\beta\delta^2 + \delta^3)\gamma^2 \\ &\quad + (\beta^2 + 6\beta\delta + 3\delta^2)\gamma^3/4 \\ &\quad + (2\beta + 3\delta)\gamma^4/16 + \gamma^5/64.\end{aligned}$$

$$\begin{aligned}\pi_1 &= 4\beta^2\delta^3 + 9\beta\delta^4 + 3\delta^5 + (4\beta\delta^3 + \delta^4)\gamma \\ &\quad + (12\beta^2\delta^2 + 28\beta\delta^3 + 20\delta^4)\gamma/4 \\ &\quad + (6\beta\delta^2 + 4\delta^3)\gamma^2 + (4\beta\delta + 6\delta^2)\gamma^3/4 \\ &\quad + (\beta + 4\delta)\gamma^4/16 + \gamma^5/64.\end{aligned}$$

$$\begin{aligned}\pi_0 &= 5\beta\delta^4 + 11\delta^5 + 5\gamma\delta^4 + (20\beta\delta^3 + 40\delta^4)\gamma/4 \\ &\quad + 10\gamma^2\delta^3 + 10\delta^2\gamma^3/4 + 5\delta\gamma^4/16 + \gamma^5/64. \qquad (29)\end{aligned}$$

The probability of correct reception is $\bar{Q} = \pi_5$, and the rate R is calculated from (9).

[14] C. W. Helstrom, "The Performance of the Scherer Code in Quaternary Symmetric Channel," Westinghouse Elec. Corp. Res. Dept. 412FF471–R2; September 1, 1959.

ACKNOWLEDGMENT

The calculations on the digital computer were programmed and supervised by Dr. H. Gordon Rice and Francis O'Meara. We wish to thank E. Scherer and Dr. R. Filipowsky of the Advanced Development Section, Westinghouse Electronics Division, who suggested the study that led to the results described here.

Linear-Recurrent Binary Error-Correcting Codes for Memoryless Channels*

WILLIAM L. KILMER†, ASSOCIATE MEMBER, IRE

Summary—**This paper concerns the analysis of recurrent-type, arity-check, error-correcting codes for memoryless, binary symetric channels. These codes are defined to consist of message equences augmented by insertions of r successive parity digits very b successive message digits. An analysis framework is stablished for the codes which consists mainly of a parity check atrix [M] and a message difference vector [N]. Within this framework, a decoding scheme is developed which renders the codes apable of correcting any set of $\leq e$ errors in m/b successive $(b + r)$-igit blocks of coded message sequence, where e is maximized ver all parity-check codes having the same redundancy ratios nd maximal lengths of dependence among their digits. An example given of a linear-recurrent code which has a lower probability f error than the best comparable block code, and several outtanding problems are discussed.**

* Received by the PGIT, February 15, 1960. This work was upported by Air Force Rome under contract No. AF 30(602)–1915.

† Elec. Engrg. Dept., Montana State College, Bozeman.

INTRODUCTION

MOST of the results to date in the theory of error-correcting codes for memoryless, binary symmetric channels have concerned block codes. These codes have almost always been of the systematic type, where n-digit code words are divided up into k information digits and $(n - k)$ parity check digits. Such codes have the advantage that they can be simple to instrument, they can be designed to have a good amount of error-correcting ability, they possess enough mathematical structure to render them amenable to analysis, and their independence between blocks gives them a decoding stability that is not easily obtainable with "recurrent-type" codes whose correct decoding at one time tends to depend upon correct decoding at all previous times.

Nevertheless, it is known that under certain circumstances essentially nonblock codes can also be designed to have easily instrumentable and peculiarly efficient error-correcting properties.[1,2] The purpose of this paper is to provide a first look into these properties for the memoryless binary symmetric channel case. This is done by first establishing a general analytical framework for a regular, parity-check type of recurrent code, and then exhibiting an example of one such code which has a greater error-correcting capability than the corresponding best block code.[3]

The Generalized Coder and Channel

The generalized coding scheme that is considered in this paper is illustrated in Fig. 1. There a random sequence of binary message digits is fed into an m-place shift register R in blocks of b digits per block. This causes corresponding blocks of b digits to be shifted out of the right end of R into the channel for transmission. After each new b-digit block is located in R, r successive parity check digits are formed in the parity check circuit P, and they in turn are moved out into the channel for transmission. The parity digits have values such that their modulo 2 sums with the corresponding message digits are 0. Thus the output of the coder consists of alternations of r successive parity digits from P, and b successive message digits from R. The code is called a *linear-recurrent code*.

The *channel* is memoryless, symmetric, and binary. Thus it is completely specified by stating that the probability of a coded message digit's being put in error is Pe, where $0 < Pe < 1/2$.

The Generalized Decoder

Hereafter, let us assume synchronous circuitry everywhere. With this understanding, the generalized decoder is given in block diagram form in Fig. 2. The system there is interested in decoding and correcting only b of the possibly-corrupted received message digits at a time. The idea behind this, roughly speaking, is to provide a decoder which is not forced to cut off from consideration any of the parity information that is available on any received message digit while it is decoding it. It is also to provide a coding system for which it is not necessary to confine all the parity information on each block of m, x_i-digits to just $m(b + r)/b$ consecutive coded message digits as in the corresponding block-code case.

Now let us make this intuitive idea precise. Suppose in Fig. 1 that a message sequence

$$X = \cdots x_{m+j}x_{m-1+j} \cdots x_{1+j} \cdots$$

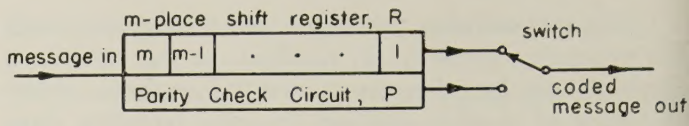


Fig. 1—Generalized coder.

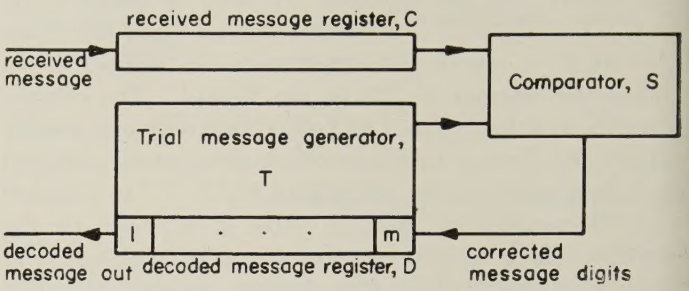


Fig. 2—Generalized decoder.

is put into the coder in the order of increasing subscripts, and that a corresponding coded message sequence

$$C(X) = \cdots c_{k+j}c_{k-1+j} \cdots c_{1+j} \cdots$$

$[k = (b + r/b)m]$ is put out of the coder. Let the noise in the channel between the coder and decoder be represented by

$$N = \cdots n_{k+j}n_{k-1+j} \cdots n_{1+j} \cdots,$$

where $n_i = 1$ if an error occurs in the ith place of $C(X)$, and $n_i = 0$ otherwise. Also let the *sum* of any two binary sequences X and Y be denoted $X + Y$, and define this sum to be the sequence

$$Z = \cdots z_{i+1}z_iz_{i-1} \cdots,$$

where z_i is the modulo 2 sum of x_i and y_i. *Then the received message in Fig. 2 can be written as* $C(X) + N$.

Suppose now that all the noise digits n_i up to and including some n_j, are 0, that at least one of $n_{j+1}, \cdots, n_{j+b}$ is 1, and that n_j is in the last of a string of r parity digits. The coded message digits that are affected by n_{j+1} and subsequent noise digits move into the decoder, a digit at a time, in blocks of $b + r$ digits per block.

At this point, recall that the decoder is interested in decoding only b, x_i-digits in $C(X) + N$ at a time. Further, note that each x_i digit can have parity digits in at most m/b successive $(b + r)$-digit blocks dependent upon it (this can be seen from Fig. 1), and that these blocks all either immediately precede or include the block of the x_i in question. Then for the case above, the decoder proceeds to investigate only the m/b, $(b + r)$-digit blocks of the received message which *end* with the digit in the c_{j+b} position. It does this by coding up for trial consideration all the m-digit sequences X_i which could possibly have gone into the coder to cause the part of its actual output sequence

$$\underbrace{\underbrace{c_{j+b} \cdots c_{j+1}}_{x_i \text{ digits}}\ \underbrace{c_j \cdots c_{j-r+1}}_{\text{parity digits}}}_{(m/b)\text{th block of } b+r \text{ digits}} \cdots \underbrace{c \cdots c}_{\text{1st block of } b+r \text{ digits}} \triangleq C(X_0). \qquad (1)$$

[1] D. W. Hagelbarger, "Recurrent codes: easily mechanized, burst-correcting, binary codes," *Bell Sys. Tech. J.*, vol. 38–4, pp. 969–984; July, 1959.

[2] W. L. Kilmer, Some Results on Best Recurrent-Type Binary Error-Correcting Codes," 1960 IRE International Convention Record, pt. 4, pp. 135–147.

[3] A. B. Fontaine, and W. W. Peterson, "Group code equivalence and optimum Codes," IRE Trans. on Information Theory, vol. IT-5, pp. 60–70; May, 1959.

bel each of the 2^m trial coded sequences, $C(X_i)$, $= 1, 2, \cdots, 2^m$).

In Fig. 2 the trial coding is accomplished in T by T's awing on the contents of D. At each point in time, D assumed to contain the m perfectly-corrected x_i digits ich just precede the x_i digits in $C(X_0)$. (The impli-tions of this assumption will be discussed later.)

Now label the subsequence,

$$\underbrace{n_{j+b} \cdots n_{j-r+1}} \cdots \underbrace{n \cdots n} \qquad (2)$$

(m/b)th block of $b + r$ noise circuits — 1st block of $b + r$ noise digits.

ich is contained in the particular N defined above, N_0. en the S part of the decoder in Fig. 2 adds each $C(X_i)$ med in T to the received sequence $C(X_0) + N_0$. The ult is a set of 2^m sequences:

$$S_i = C(X_i) + C(X_0) + N_0, \qquad (i = 1, 2, \cdots, 2^m).$$

ce the coder and decoder are linear, we have that

$$(X_i) + C(X_0) = C(X_i + X_0), \qquad (i = 1, 2, \cdots, 2^m).$$

erefore, when $X_i = X_0$, $S_i = N_0$, and otherwise $= N_0 + C(X_i + X_0)$.

At this point let us suppose we have a code such that r every i for which $(X_i + X_0)$ contains at least one 1 one of its first θ places, $\theta \leqslant m$, $C(X_i + X_0)$ contains minimum of $(2e + 1)$ 1's over-all. Suppose also that e decoder selects at each "stage" of decoding the st θ digits of an X_i for which the corresponding $(X_i + X_0) + N_0]$ contains a minimum number of 1's. en over successive stages of decoding, the code in estion provides e-error correction within any set of $/b)$ consecutive $(b + r)$-digit blocks of coded message gits. It is clear that this provision is on a maximum-elihood-detection basis.

At first it might seem that for any given code, the aller θ is, the fewer X_i there are that differ from X_0 one of the first θ places, and hence the fewer $C(X_i + X_0)$ ere are to consider in establishing the e in $(2e + 1)$. ut the fact is that nothing is gained by reducing θ to ything less than b. The reason is simply that for $\theta < b$, the parity information that pertains to the first b gits of X_0 is used in the selection of the θ digits. Hence ecting only θ, x_i-digits at each stage of decoding does t allow the introduction of more parity information on e remaining x_i's of the block during the next stage of coding. So from now on, we will assume that $\theta = b$.[4]

Quasi-Best Codes

A logical question to ask at this point is: how does one ect parity relations for P in Fig. 1 so as to maximize e a function of m, r, and b? This question may *not* be uivalent to asking which codes provide minimum prob-ilities of error (*i.e.*, are "optimum"). For consider the (block-type) group code analog: there it is not always true that the optimum codes are those which have the maximum e's,[5] neither is it always true that the codes which are optimum for small p_e are also optimum for large p_e (the dividing line between small and large p_e is usually about 0.3).[6] But these statements are *usually* true. On this basis, it seems justifiable to define *quasi-best linear-recurrent codes* as those which have maximum e's (e either an integer or an integer divided by 2).

At this point we omit a rather obvious proof, involving simple rearrangements of parity and message digits, and simply state that a *quasi-best linear-recurrent code is also one which can correct any set of e errors in $(b + r)m/b$ successive coded information digits, where e is now maximum over all equally-redundant parity check codes whose parity digits are not dependent upon any more than m successive message digits.* It is important to note in this context that if for some code $m = b$, and P is not fixed but varies periodically with period r, the code in question is simply a usual block type of parity check code. *Thus the set of all linear-recurrent codes properly contains the set of all parity check block codes.*

To proceed with the above definition of quasi-best code as our criterion, we now desire codes which meet the following specifications: for given m, r, and b, the minimum number of 1's in $C(X_i) + C(X_0)$, taken over all X_i for which $(X_i + X_0)$ contains at least one 1 in its first b places, is an absolute maximum. In other words, for all X_i in question, if $(X_i + X_0)$ contains q 1's, the parity digits in $C(X_i)$ differ from those in $C(X_0)$ in at least $j - q$ places, for maximum possible j.

In order to pursue the desired codes in mathematical terms, let us define an m-digit *parity check vector*

$$p^j = p_m^j p_{m-1}^j \cdots p_1^j$$

as a representation of the jth parity check relation in P of Fig. 1 as follows: if the jth parity digit being formed is dependent upon the digit in the ith place of R, $p_i^j = 1$; otherwise $p_i^j = 0$. The significance of the j here is that in the general case the parity relation in P varies over some period.

Next let us formulate an expression which, for any particular code as defined by a sequence of p^j's, exhibits the places in which the parity digits of a $C(X_i)$ differ from those in $C(X_0)$. To do this, define $C^*(X_i + X_0) \triangleq C_i^*$ to be $C(X_i + X_0)$ with everything but parity-position digits deleted; and define $N_i^* = (X_i + X_0)$. Then the 1's in C_i^* indicate just where the parity digits in $C(X_i)$ differ from those in $C(X_0)$.

The first digit in C_i^* is given by the modulo 2 value of the ordinary matrix product,

$$[p_m^{k+1} p_{m-1}^{k+1} \cdots p_{m-b+1}^{k+1}] \begin{bmatrix} n_{i_b}^* \\ \vdots \\ n_{i_1}^* \end{bmatrix}, \qquad (3)$$

[4] The idea here of decoding by stages is related to that contained J. M. Wozencraft, "Sequential Decoding for Reliable Com-unication," Elec. Engrg. Dept., Mass. Inst. Tech., Cambridge, ass.

[5] D. Slepian, "A class of binary signaling alphabets," *Bell Sys. Tech. J.*, vol. 35, pp. 203–234; Sec. 1.10; January, 1956.

[6] Fontaine and Peterson, *op. cit.*, pp. 67–68.

where p^{k+1} is the parity relation used to establish the first c_i in (1), and $n^*_{i_1}, \cdots, n^*_{i_b}$ are the 1st to the bth digits respectively in N^*_i. The second digit in C^*_i is given similarly by the modulo 2 value of

$$[p_m^{k+2} p_{m-1}^{k+2} \cdots p_{m-b+1}^{k+2}] \begin{bmatrix} n^*_{i_b} \\ \vdots \\ n^*_{i_1} \end{bmatrix}. \tag{4}$$

Continuing in this manner, the first r digits in C^*_i are given by the modulo 2 values, from bottom to top in the order of their occurrence, of the numbers in the column sequence equal to

$$\begin{bmatrix} p_m^{k+r} & \cdots & p_{m-b+1}^{k+r} \\ \vdots & & \\ p_m^{k+1} & \cdots & p_{m-b+1}^{k+1} \end{bmatrix} \begin{bmatrix} n^*_{i_b} \\ \vdots \\ n^*_{i_1} \end{bmatrix}. \tag{5}$$

Finally, since the last digit in C^*_i is the last parity digit in the (m/b)th block of digits in (1), all the digits in C^*_i are given by the modulo 2 values, from bottom to top in the order of their occurrence, of the numbers in the column sequence equal to

$$\left[\begin{array}{ccc|c|ccc} p_m^{k+mr/b} & \cdots & p_{m-b+1}^{k+mr/b} & \cdots & p_b^{k+mr/b} & \cdots & p_1^{k+mr/b} \\ \vdots & & & & \vdots & & \\ p_m^{k+(m-1/b)r+1} & \cdots & p_{m-b+1}^{k+(m-1/b)r+1} & & p_b^{k+(m/b-1)r+1} & \cdots & p_1^{k+(m/b-1)r+1} \\ \hline & & & & & & \\ & & & & & & \\ \hline & & p_m^{k+2r} & \cdots & p_{m-b+1}^{k+2r} & p_{m-1}^{k+2r} & \cdots & p_{m-2b+1}^{k+2r} \\ & & \vdots & & & \vdots & & \\ & & p_m^{k+r+1} & \cdots & p_{m-b+1}^{k+r+1} & p_{m-b}^{k+r+1} & \cdots & p_{m-2b+1}^{k+r+1} \\ \hline & & & & & p_m^{k+r} & \cdots & p_{m-b+1}^{k+r} \\ & & & & & \vdots & & \\ & & & & & p_m^{k+1} & \cdots & p_{m-b+1}^{k+1} \end{array}\right] \begin{bmatrix} n^*_{i_m} \\ \vdots \\ n^*_{i_{m-b+1}} \\ \vdots \\ n^*_{i_b} \\ \vdots \\ n^*_{i_1} \end{bmatrix}, \tag{6}$$

$$\triangleq \begin{bmatrix} M_{1,1} & M_{1,2} & \cdots & M_{1,m/b} \\ & M_{2,2} & \cdots & M_{2,m/b} \\ & & \ddots & \vdots \\ & & M_{m/b-1,m/b-1} & M_{m/b-1,m/b} \\ & & & M_{m/b,m/b} \end{bmatrix} [N^{**}_i], \tag{7}$$

$$\triangleq [M][N^{**}_i] \triangleq [C^{**}_i] \tag{8}$$

where $[C^{**}_i]$ and $[N^{**}_i]$ are just the transposes of the C^*_i and N^*_i sequences, respectively, in reverse order from left to right, and all the elements below the solid lines in (6) and (7) are 0. Note that in (6), the modulo 2 product of the bottom row of $[M]$ and $[N^{**}_i]$ is just (4) above; the product of the bottom row of (7) [*i.e.*, the bottom r rows of $[M]$ in (6)] and $[N^{**}_i]$ is just (5) above; etc. Also note that the validity of (3) to (8) depends on the assumption that the previously decoded $(m/b - 1)$ blocks contain no error.

Now a most important feature of (6), which is just the upside-down transpose of C^*_i, is that it is equal to the sum of those columns in $[M]$ for which there correspond 1's in the same rows of $[N^{**}_i]$, counting from the bottom up. This is because postmultiplying $[M]$ by a column vector amounts to forming the linear combination of its columns specified by the postmultiplier. Thus the nature of C^*_i and hence the value of e is essentially established by the columnar sum properties of $[M]$. This fact is basic to the study of linear recurrent codes.

Now we are in a position to discuss exactly what it is that we desire: *For given m, r, and b, we want to select a periodic sequence $p^1, p^2, \cdots p^z$, with z finite and such that for k in* (6) *equal to* 1, 2, $\cdots$, *or z; $(k + 1)$ in* (6) *equal to* 2, 3, $\cdots$, z, *or* 1, *respectively, etc; the sum of any subset of q columns of $[M]$ contains at least $(j - p)$* 1's, *for maximum possible j.* In case it is required that the parity check vector be fixed at p, it is sufficient to select only the top row of $[M]$ to completely specify it. Having chosen the required list of p's, the complete logical specification of the coder in Fig. 1 follows immediately from the definition of a parity check vector.

Relation to Block Coding Theory

At this point there remains the problem of specifying a procedure for constructing p's according to the criterion just given. It is interesting to note that this problem is almost identical to a corresponding one, given by Sacks,[7]

[7] G. E. Sacks, "Multiple error correction by means of parity checks," IRE Trans. on Information Theory, vol. IT-4, pp. 145–147; December, 1958.

r constructing e-error-correcting block-type group codes. cks' problem, which is still unsolved in any satisfactory nse, is to choose $(n - r)$ binary sequences (the code aracteristics) to serve as the top $(n - r)$ rows in the by r matrix such that no subset of $\leqslant 2e$ of the rows in

$$\begin{array}{cc} & \text{columns} \\ & 1 \quad 2 \quad \cdots \quad r \\ \text{rows} \begin{array}{c} 1 \\ 2 \\ \vdots \\ n-r \\ n-r+1 \\ \vdots \\ n \end{array} & \left[\begin{array}{c} \text{rows consisting of the } (n-r) \text{ code characteristics} \\ \hline r \text{ by } r \text{ unit submatrix} \end{array}\right] \end{array} \tag{9}$$

e linearly dependent. Because of the unit submatrix in), this problem is equivalent to the one of choosing $- r)$ rows such that no sum of any q of these rows ntains fewer than $(2e + 1 - q)$ 1's in it. Clearly, this oblem is almost the "transpose" of the general one at is stated above for linear-recurrent codes (in terms the columnar sum properties of the $[M]$ matrices for ese codes). The only differences are that: 1) in the case linear-recurrent codes all the below-diagonal elements $[M]$ have to be 0, whereas in the upper submatrix of), full-length sequences are allowed; and 2) for linear- current codes at least one of the rightmost b columns $[M]$ has to be included in every one of $[M]$'s relevant lumn sums, whereas no such restriction is present in e group code problem.

Example of a Best Linear-Recurrent Code

An example will now be given of a quasi-best linear- current code which is actually best in the following nse: over a memoryless, binary symmetric channel this de provides the lowest probability of error that any ear (parity check) code whose interdigit dependence tends over a maximum of 12 digits possibly can. This mains true as long as no errors are made in the decoding eration.

The code is defined by the parity check vector $p = 1001$, so the $[M]$ for the code is

$$\begin{bmatrix} 1 & 1 & 1 & 0 & 0 & 1 \\ & 1 & 1 & 1 & 0 & 0 \\ & & 1 & 1 & 1 & 0 \\ & & & 1 & 1 & 1 \\ & & & & 1 & 1 \\ & & & & & 1 \end{bmatrix}.$$

This matrix has the property that every sum of q of columns that includes its rightmost column contains least $(5 - q) = (2e + 1 - q)$ 1's. Hence, according the decoding scheme given above, the code has the ability to correct all single and double errors which can occur in any set of $(m/b) = 6$ consecutive blocks of coded message digits, where there are $(b + r) = 2$ digits per block. This compares favorably with the best (12, 6) block code[8] which can only correct all single errors, 50 of the 66 possible double errors, and 1 of the 286 possible triple errors that can occur in any block of 12 digits[9].

The decoder in this example decides on $b = 1$ message digits per stage, where $2^6 = 64$, 6-digit X_i's are coded up for trial consideration at each stage. It is probably no surprise that the 64 X_i's here do not imply an excessive amount of equipment in the decoder. For the T circuit can consist essentially of: 1) a 6-digit linear feedback shift register capable of generating a $(2^6 - 1)$-length binary sequence,[10] which sequence automatically contains every 6-digit binary sequence as a consecutive-digit subsequence; and 2) a single coder of the type shown in Fig. 1 to code up the output of the shift register. S in Fig. 2 can consist simply of a 12-digit sequence comparator (*i.e.*, a 2-digit modulo 2 adder and a count-to-3 counter), and a storage register for storing at any time t the best trial X_i found from the beginning of the decoding stage occurring at t until t.

Some Outstanding Problems

Up until this point it has been assumed that mistakes never occur in the decoding process. This assumption has enabled us to avail the following idea, which essentially restates what has gone before, but from a slightly different point of view: At each stage of decoding, the block of b message digits being decoded has first of all r parity digits generated which are (effectively) dependent solely upon it; then the block being decoded has r more parity digits generated which are (effectively) dependent only upon it and the succeeding block of message digits, where the succeeding block is chosen so as to provide a best match to the received coded message, and so on, until finally the block being decoded has an (m/b)th set of r parity digits generated which are (effectively) dependent upon it and the $(m/b - 1)$ succeeding blocks of message digits, where each of these succeeding blocks is chosen so as to provide a best match to the received coded message. Thus, under the assumption that no mistakes ever occur in the decoding process, this process amounts, essentially, to the successive generation and subsequent deciphering of an aggregate of redundancy patterns (the blocks of r parity digits), where this aggregate is always "*weighted*" most heavily toward that end of the message which is currently being decoded.

With this scheme of decoding, if the results previous to any stage leave incorrect digits in D, some very disrupting things can happen. In order to mathematize them, suppose that at some stage of decoding the mistakes in D

[8] Containing 6 message digits and 6 parity digits per 12-digit block.

[9] Fontaine and Peterson, *op. cit.*, p. 68.

[10] B. Elspas, "The theory of autonomous linear sequential networks," IRE Trans. on Circuit Theory, vol. CT-6, pp. 45–60; March, 1959.

of Fig. 2 are indicated by the column vector

$$\begin{bmatrix} d_m \\ \vdots \\ d_1 \end{bmatrix} \triangleq [D^*], \tag{10}$$

where $d_i = 1$ if the digit in the ith place of D is incorrect, and $d_i = 0$ otherwise. Then instead of (8), we have, after the *manner* of (8),

$$[C_i^{**}] = [M][N^{**}] + [M'][D^*], \tag{11}$$

where

$$[M'] = \begin{bmatrix}
 & & & & & & & & \\
p_b^{k+(m/b-1)r} & \cdots & p_1^{k+(m/b-1)r} & & & & & & \\
\vdots & & \vdots & & & & & & \\
p_b^{k+(m/b-2)r+1} & \cdots & p_1^{k+(m/b-2)r+1} & & & & & & \\
 & & & \ddots & & & & & \\
p_{m-2b}^{k+2r} & \cdots & p_{m-3b+1}^{k+2r} & & p_b^{k+2r} & \cdots & p_1^{k+2r} & & \\
\vdots & & \vdots & \cdots & \vdots & & \vdots & & \\
p_{m-2b}^{k+r+1} & \cdots & p_{m-3b+1}^{k+r+1} & & p_b^{k+r+1} & \cdots & p_1^{k+r+1} & & \\
p_{m-b}^{k+r} & \cdots & p_{m-2b+1}^{k+r} & & p_{2b}^{k+r} & \cdots & p_{b+1}^{k+r} & p_b^{k+r} \cdots p_1^{k+r} & \\
\vdots & & \vdots & \cdots & \vdots & & \vdots & \vdots & \\
p_{m-b}^{k+1} & \cdots & p_{m-2b+1}^{k+1} & & p_{2b}^{k+1} & \cdots & p_{b+1}^{k+1} & p_b^{k+1} \cdots p_1^{k+1} &
\end{bmatrix}$$

with all the elements above the solid line 0, and $[D^*]$ is as defined in (10). The validity of (11) can be seen by recalling that when the first r parity digits are formed for some trial X_i at any stage of decoding, the first b digits of that X_i have already moved into the leftmost b places of T's coder register; and so on, until at the end of the X_i trial, all m of the X_i digits have moved into T's coder register.

Now if $[D^*]$ does not contain all 0's, the weighted-redundancy idea stated above breaks down almost completely. This is reflected in the analysis by the fact that in this case *the following new criterion* has to be substituted into the analysis scheme: The $[C_i^{**}]$'s, for all i for which $(X_i + X_0)$ contains at least one 1 in the first b places, must all differ from each other in at least $(2e + 1 - q)$ places regardless of what $[D^*]$ happens to be. This must be so if e-error-correction is to be obtained over every (m/b) successive $(b + r)$-digit blocks of the coded message sequence.

Thus far, this new criterion has seemed especially intractable—at least the author has not yet been able to do anything very satisfactory with it. The following is all that can be reported: If a decoding mistake occurs with a quasi-best code, in all probability the future decoding operations will be hopelessly garbled, but the probability is not 0 that the decoder will eventually get back on the right track.[11] If garbling does occur, it is highly probable that no trial $C(X_i)$'s will perfectly match the received $(C(X_0) + N_0)$'s for a very long time, even though many or most of these N_0's may be 0. The one bright aspect here is that for low-noise channels, this fact might serve as a kind of *infinite-degree error-detection criterion*!

Another outstanding problem is: how are good, quasi-best, or best codes constructed for general m, r, and b? It should be mentioned here that the example of the previous section was not obtained in the presence of any such generality. Rather, it was one of a set of fixed-parity-check codes that were derived for all $m \leqslant 6$ and for some $m > 6$ by exhaustive and trial-and-error methods. All that can be said from the results is that apparently a tight description of the summability properties (in the number-of-1's sense) of nonperiodic, nonrandom, binary sequences is needed before further general progress can be made. The author is reminded, in this regard, of the work of Golomb,[12] Zierler,[13] and Rothstein,[14] but would at present make no claims in these directions.

[11] For N_0 could be 0's in the message-digit positions, and 1's in just those parity-digit positions where corresponding 1's were indicated in $[M']$ $[D^*]$; and this could happen enough times in succession to allow enough perfect matches between correct trial $C(X_i)$'s and received $(C(X_0) + N_0)$'s to reduce $[D^*]$ to all 0's as required.

[12] S. W. Golomb, "Sequences With Randomness Properties," Glenn L. Martin Co., Baltimore, Md., Final Rept. on Contract No. SC–54–33611; advance copy dated June 14, 1955.

[13] N. Zierler, "Several Binary Sequence Generators," Lincoln Lab., Mass. Inst. Tech., Lexington, Mass., Tech. Rept. No. 95; September, 12, 1955.

[14] J. Rothstein, "Analysis of binary time series in periodic functions," IRE TRANS. ON ELECTRONIC COMPUTERS, vol. EC-8, p. 229; June, 1959.

Summary and Conclusions

In this paper an analysis framework is established for ear-recurrent codes meant for use over memoryless, nary symmetric channels. Within the context of the amework, consisting mainly of the matrices $[M]$ and $[^{**}_{i}]$, a criterion is given for constructing "quasi-best" des capable of correcting any set of $\leqslant e$ errors in (m/b) ccessive $(b + r)$-digit blocks of a coded message se- ence, for maximum possible e. This criterion is compared a similar one given by Sacks for corresponding best ock codes. An example is given of a linear-recurrent de which has a lower probability of error than the corresponding best block code. Finally, several problems are suggested, mainly with a view towards finding more efficient codes for practical uses, and a general way to approach channel capacity at a faster rate (with respect to increasing maximum length of dependence among coded message digits) than is possible with ordinary block codes. The latter might reasonably be expected, for the "weighted-redundancy" and "increased-effective-block-length" effects that are achieved by linear recurrent coding would seem to suggest that as m increases, linear-recurrent codes get better and better than their counterpart block codes.

Probability Density Functions for Correlators with Noisy Reference Signals*

G. M. ROE† and G. M. WHITE†, member, ire

Summary—**Recently, correlation functions have had to be considered where both the reference waveform, which is usually the esired signal, and the input waveform are masked by different mples of additive noise. In this article, we derive the probability ensity function for the random variable β where**

$$\beta = \sum_{i=1}^{k} (As_{i,x} + N_{i,x})(Bs_{i,y} + N_{i,y}).$$

he $s_{i,x}$ and $s_{i,y}$ are the signal components, and $N_{i,x}$ and $N_{i,y}$ are mples of Gaussian noise.

Exact expressions involving Bessel and Whittaker functions are ven for several cases. Asymptotic expressions allow $W(\beta)$ to be otted when these exact expressions cannot be obtained or conniently evaluated.

Introduction

CROSS-CORRELATION techniques of comparing two waveforms have been playing an increasingly important role in communication or radar systems. n typical models one of the waveforms used in the cross-orrelation computation is a clean signal unperturbed by oise. This can be considered as the reference waveform. he other waveform, which can be called the input wave-orm, is usually either Gaussian noise or the same signal, erturbed by additive Gaussian noise. In these cases, the utput of the correlator is Gaussian distributed and has zero or nonzero mean, depending upon the absence or resence of the signal.

If the functions to be cross-correlated are bandwidth limited and the integration is performed for only a finite length of time, the integral normally used can be approximated by a sum of products.[1] The terms forming these products are amplitude samples of the two waveforms at the Nyquist intervals. The output of this summing circuit, if the reference is again a clean signal, is Gaussian distributed, having a variance of $k\sigma^2$, where k is the number of taps and σ is the standard deviation of the noise distribution. If a signal is present, the mean is determined by the auto-correlation function of the signal. If there is exact alignment between the components, the mean of the output distribution is then

$$\sum_{i=1}^{k} s_i^2,$$

where s_i is the signal component and k is the number of taps.

In this article, we consider a more complicated situation where the reference waveform is no longer a clean signal, but is corrupted by additive Gaussian noise. This problem originated from a theoretical analysis of an adaptive waveform recognizer,[2] and is similar, when there are no signal components present, to the problem proposed by

* Received by the PGIT, July 1, 1960.
† General Electric Res. Lab., Schenectady, N. Y.

[1] P. M. Woodward, "Probability and Information Theory With Applications to Radar," McGraw-Hill Book Co., Inc., New York, N. Y.; 1953.
[2] C. V. Jakowatz, R. L. Shuey, and G. M. White, "Adaptive Waveform Recognition," presented at the Symp. on Information Theory, London, Eng.; Sept. 1, 1960.

D. G. Lampard.[3] We assume here that both the input and reference waveforms are bandlimited functions and that the correlation can be performed by taking the sum of products. Under these conditions, the output of the correlator is no longer Gaussian. For special cases, we have found exact expressions for the desired density functions which involve series of Bessel functions or series of Whittaker functions. These series are not particularly efficient for purposes of numerical computations, but alternative approximations based on a saddle point integration are sufficiently accurate for most purposes.

The Problem

In many communication systems, it is desirable to compute the integral

$$2W\beta(\tau) = \int_0^T x(t)y(t + \tau)\, dt \tag{1}$$

where $x(t)$ and $y(t)$ are two waveforms.

If $x(t)$ and $y(t)$ are bandwidth-limited functions where the bandwidth extends from $-W$ to $+W$, then the integral given in (1) can be approximated by

$$\beta = \sum_{i=1}^{k} X_i Y_i \tag{2}$$

where X_i and Y_i are k samples at the Nyquist intervals corresponding to T and W.

A correlator that can perform this cross-correlation function is shown in Fig. 1. The taps of the delay line are placed at the Nyquist intervals, and there are k of them. The inputs $x(t)$ and $y(t)$ each may be noise or signal and noise. The noise in input $x(t)$ is usually independent of the noise in $y(t)$. Furthermore, since the taps are located at Nyquist intervals, the values of the noise samples are independent of each other.

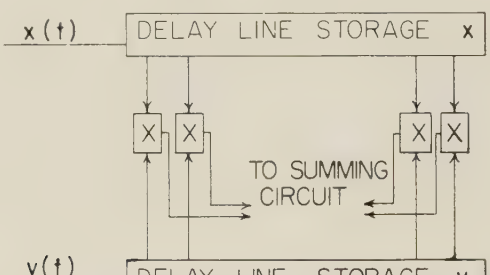


Fig. 1—Correlator.

All of the special cases of interest may be included in the general case where the output of the correlator is written as

$$\beta = \sum_{i=1}^{k} X_i Y_i = \sum_{i=1}^{k} (As_{i,x} + N_{i,x})(Bs_{i,y} + N_{i,y}). \tag{3}$$

[3] D. G. Lampard, "The probability distribution for the filtered output of a multiplier whose inputs are correlated, stationary, Gaussian time-series," IRE Trans. on Information Theory, vol. IT-2, pp. 4–11; March, 1956.

The $N_{i,x}$ and $N_{i,y}$ are Gaussian distributed, and

$$\langle N_{i,x}N_{j,x}\rangle = \sigma_x^2\, \delta_{ij}, \tag{4}$$

$$\langle N_{i,y}N_{j,y}\rangle = \sigma_y^2\, \delta_{ij}, \tag{5}$$

and

$$\langle N_{i,x}\, N_{j,y}\rangle = \rho\sigma_x\sigma_y\, \delta_{ij}. \tag{6}$$

Thus the case where the signals in both channels are aligned can be obtained by taking $\rho = 0$, $s_{i,x} = s_{i,y}$ and $A = B = 1$; while letting $\rho = 0$, and $A = 1$, $B = 0$ yields the case of signal plus noise in the reference channel and pure noise in the input channel. If the signals are not aligned, but the signal components are shifted m Nyquist intervals in the input channel, then $s_{i,y} = s_{i,x+m}$.

It is convenient to denote the normalized output of the correlator by ϕ:

$$\phi = \frac{\beta}{\sigma^2} \tag{7}$$

and

$$\sigma^2 = \sigma_x\sigma_y. \tag{8}$$

The density functions we desire are for the normalized ϕ. Furthermore, the B in (3) can be adjusted so that

$$\sum_{i=1}^{k} s_{i,x}^2 = \sum_{i=1}^{k} s_{i,y}^2. \tag{9}$$

R, the measure of the signal-to-noise ratio, is defined as

$$R = \frac{1}{\sigma^2}\sum_{i=1}^{k} s_{i,x}^2 = \frac{1}{\sigma^2}\sum_{i=1}^{k} s_{i,y}^2. \tag{10}$$

If $\sigma_x = \sigma_y$, then

$$R = \frac{2E}{N_0} \tag{11}$$

where E is the signal energy and N_0 is the mean noise power per unit bandwidth. This definition is the equivalent to Woodward's R.[1] It is also convenient to define

$$\gamma = \frac{R}{k(1 - \rho^2)}, \tag{12}$$

$$\eta = \frac{\phi}{k(1 - \rho^2)} = \frac{\beta}{k\sigma^2(1 - \rho^2)}. \tag{13}$$

Derivation of the Density Function

In order to derive the density function $W(\phi)$, we start with the joint density function $W_2(x_i, y_i)$ for the variables

$$\begin{aligned} x_i &= \frac{X_i}{\sigma_x} = \frac{As_{i,x}}{\sigma_x} + \frac{N_{i,x}}{\sigma_x} = as_{i,x} + n_{i,x}, \\ y_i &= \frac{Y_i}{\sigma_y} = \frac{Bs_{i,y}}{\sigma_y} + \frac{N_{i,y}}{\sigma_y} = bs_{i,y} + n_{i,y} \end{aligned} \tag{14}$$

where we consider an ensemble of correlators of the type shown in Fig. 1. The terms $as_{i,x}$ and $bs_{i,y}$ are treated as constants of the ensemble, and the noise terms are assumed to be Gaussian and $n_{i,x}$ correlated only with $n_{i,y}$. The

int density function is therefore

$$W_2(x_i, y_i) = \frac{1}{2\pi(1-\rho^2)^{1/2}} \exp - \frac{1}{2(1-\rho^2)} \cdot\{(x_i - as_{i,x})^2 - 2\rho(x_i - as_{i,x}) \cdot(y_i - bs_{i,y}) + (y_i - bs_{i,y})^2\}. \quad (15)$$

Ve now change from the pair (x_i, y_i) to the pair (x_i, ϕ_i) here $\phi_i = x_i y_i$. Since the Jacobian for the transformation is $J = \mid 1/x_i \mid$, the new joint distribution is

$$W_2(x_i, \phi_i) = \frac{1}{2\pi \mid x_i \mid (1-\rho^2)^{1/2}} \exp - \frac{1}{2(1-\rho^2)} \cdot\{(x_i - as_{i,x})^2 - 2\rho(x_i - as_{i,x}) \cdot(\phi_i/x_i - bs_{i,y}) + (\phi_i/x_i - bs_{i,y})^2\}. \quad (16)$$

he Fourier transform relative to ϕ_i is

$$\bar{W}_2(x_i, \xi_i) = \int_{-\infty}^{\infty} W_2(x_i, \phi_i) e^{i\xi_i\phi_i}\, d\phi_i \quad (17)$$

$$\bar{W}_2(x_i, \xi_i) = \frac{1}{\sqrt{2\pi}} \exp - \tfrac{1}{2}\{x_i^2[1 - 2i\xi_i\rho + \xi_i^2(1-\rho^2)] - 2x_i[as_{i,x}(1 - i\rho\xi_i) + bs_{i,y}] + a^2 s_{i,x}^2\}. \quad (18)$$

he characteristic function for ϕ_i may now be found by ntegrating this result over all x_i:

$$\bar{W}_1(\xi_i) = \frac{1}{\sqrt{1 - 2i\xi_i\rho + \xi_i^2(1-\rho^2)}} \exp \tfrac{1}{2}\left\{\frac{2i\xi_i abs_{i,x}s_{i,y} - \xi_i^2[b^2 s_{iy}^2 - 2\rho abs_{ix}s_{iy} + a^2 s_{ix}^2]}{1 - 2i\xi_i\rho + \xi_i^2(1-\rho^2)}\right\}. \quad (19)$$

ince the taps have been spaced at Nyquist intervals, he ϕ_i are independent and the characteristic function or the sum $\phi = \sum_{i=1}^{k} \phi_i$ is the product of the k characteristic functions $\bar{W}_1(\xi_i)$ or

$$\bar{W}(\xi) = [1 - 2\rho i\xi + \xi^2(1-\rho^2)]^{k/2} \cdot\exp - \sigma^2 R\, \frac{\xi^2[\tfrac{1}{2}(a^2+b^2) - \rho abr] - i\xi abr}{1 - 2\rho i\xi + \xi^2(1-\rho^2)} \quad (20)$$

here

$$r = \sum_{i=1}^{k} s_{i,x}s_{i,y} \Big/ \sum_{i=1}^{k} s_{i,x}^2.$$

we substitute

$$\xi = \frac{i\rho}{i - \rho^2} + \frac{z}{1-\rho^2},$$

then

$$W(\phi) = \frac{1}{2\pi}\int_{-\infty}^{\infty} \bar{W}(\xi) e^{-i\xi\phi}\, d\xi, \quad (21)$$

$$= \frac{L(\phi, \rho)}{2\pi} \int_{-\infty}^{\infty} (1+z^2)^{-k/2} \cdot\exp - \frac{R}{1-\rho^2}\frac{hz^2 - izc}{1+z^2} \exp - \frac{iz\phi}{1-\rho^2}\, dz \quad (22)$$

where

$$L = L(\phi, \rho) = (1-\rho^2)^{(k/2)-1} \cdot\exp \frac{\rho}{1-\rho^2}\left[\phi - R\sigma^2\left\{abr - \frac{\rho}{2}(a^2+b^2)\right\}\right], \quad (23)$$

$$c = \sigma^2[abr(1+\rho^2) - \rho(a^2+b^2)] = ABr(1+\rho^2) - \rho\left(\frac{A^2\sigma^2}{\sigma_x^2} + \frac{B^2\sigma^2}{\sigma_y^2}\right), \quad (24)$$

$$h = \sigma^2[\tfrac{1}{2}(a^2+b^2)(1+\rho^2) - 2\rho abr] = \frac{\sigma^2}{2}\left(\frac{A^2}{\sigma_x^2} + \frac{B^2}{\sigma_y^2}\right)(1+\rho^2) - 2\rho ABr. \quad (25)$$

The characteristic function in (20) can also be developed as a particular case of already established results. Middleton[4] presents the derivation of the characteristic function for the quadratic form

$$x = \underline{Y}^T \underline{J}\, \underline{Y} \quad (26)$$

where $\underline{J}$ is a symmetric matrix and $\underline{Y}$ a column vector of random components. Our variable β can be put in this form by letting the first k components of $\underline{Y}$ represent the quantities $\underline{X}_i$ and the second k components of $\underline{Y}$ represent the quantities $\underline{Y}_i$ and by choosing $\underline{J}$ of the form

$$J = \begin{bmatrix} 0 & \underline{I} \\ \underline{I} & 0 \end{bmatrix}$$

where $\underline{I}$ is the identity matrix. For the limited correlations permitted by (4)–(6), the covariance matrix for the $\underline{Y}_i$ has only a single paired set of off-diagonal terms, and the required matrix inversions can be written down directly.

The Formal Expression

Exact expressions for the integral in (22) in terms of known functions may be obtained for three cases of primary interest, as follows:

Case n − n: If there is no signal in either channel, then $A = B = c = h = 0$, and

$$W_{nn}(\phi) = \frac{L(\phi, \rho)}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-(iz\phi)/(1-\rho^2)}}{(1+z^2)^{k/2}}\, dz. \quad (27)$$

[4] D. Middleton, "An Introduction to Statistical Communication Theory," McGraw-Hill Book Co. Inc., New York, N. Y., sec. 17.2–1, p. 738; 1960.

The integral is a Bessel function of imaginary argument,[5] and we have

$$W_{nn}(\phi) = \frac{1}{\sqrt{\pi}\,\Gamma\left(\frac{k}{2}\right)(1-\rho^2)^{\frac{1}{2}}}\left(\frac{|\phi|}{2}\right)^{(k-1)/2} \cdot \exp\left(\frac{\rho}{1-\rho^2}|\phi|\right) K_{(k-1)/2}\left\{\frac{|\phi|}{1-\rho^2}\right\}. \quad (28)$$

This is the same result that Wishart and Bartlett[6] obtained when they considered a similar sum of products.

This same result can also be obtained from Lampard's[3] paper if one considers that the impulse response of the post multiplier filter is the sum of k delta functions where the arguments of these delta functions are zero at the Nyquist intervals. Since the noise is bandlimited, the correlation functions in (53) of his paper become (with $\operatorname{sinc} x = \sin \pi x/\pi x$)

$$\psi_{11}(t) = \sigma_x^2 \operatorname{sinc} 2Wt,$$

$$\psi_{12} = \rho\sigma_x\sigma_y \operatorname{sinc} 2Wt,$$

$$\psi_{22} = \sigma_y^2 \operatorname{sinc} 2Wt.$$

Case $n - s$: If there is signal plus noise in one channel, and noise only in the other, ($A = 1$, $B = 0$) and $\rho = 0$; then

$$L(\phi, 0) = 1,$$

$$c = 0,$$

$$h = \frac{\sigma^2}{2\sigma_x^2}.$$

If also $\sigma_x = \sigma_y$, then

$$W_{ns}(\phi) = \frac{1}{2\pi}\int_{-\infty}^{\infty}(1+z^2)^{-k/2} \cdot \exp\left\{-\frac{R}{2} - iz\phi\right\} \exp + \frac{R}{2(1+z^2)}\,dz. \quad (29)$$

Expansion of the second exponential factor yields a series of integrals similar to (27), and integration term by term gives

$$W_{ns}(\phi) = \frac{1}{\sqrt{\pi}}e^{-R/2} \cdot \sum_{j=0}^{\infty}\frac{1}{j!\,\Gamma\left(\frac{k}{2}+j\right)}\left(\frac{R}{2}\right)^j\left(\frac{|\phi|}{2}\right)^{(k-1+2j)/2} K_{(k-1+2j)/2}(|\phi|). \quad (30)$$

Case $s - s$: If there is signal plus noise in both channels, and the signals are aligned, then ($A = B = 1$) and $s_{ix} = s_{iy}$. Also, as in case $n - s$, $\sigma_x = \sigma_y$ and $\rho = 0$ for the conditions

$$c = 1,$$

$$h = 1,$$

and the density function becomes

$$W_{ss}(\phi) = \frac{1}{2\pi}\int_{-\infty}^{\infty}(1+z^2)^{-k/2}e^{-iz\phi-R}e^{R/(1-iz)}\,dz. \quad (31)$$

After expansion of the second exponential term, integration term by term yields a series of Whittaker functions.[7] For $\phi > 0$,

$$W_{ss}(\phi) = \tfrac{1}{2}e^{-R}\sum_{j=0}^{\infty}\frac{R^j}{j!\,\Gamma\left(\frac{k}{2}+j\right)}\left(\frac{\phi}{2}\right)^{(k+j-2)/2} W_{j/2,(k+j-1)/2}(2\phi), \quad (32)$$

and for $\phi < 0$,

$$W_{ss}(\phi) = \tfrac{1}{2}e^{-R} \cdot \sum_{j=0}^{\infty}\frac{R^j}{j!\,\Gamma\left(\frac{k}{2}\right)}\left(\frac{|\phi|}{2}\right)^{(k+j-2)/2} W_{-j/2,(k+j-1)/2}(2|\phi|). \quad (33)$$

An alternate form for these series can be obtained by expressing the Whittaker functions in terms of the hypergeometric functions ${}_2F_0$.[8]

The density function that Marcum[9] derives for composite pulses of signal-plus-noise minus noise can be shown to be special cases of the functions given by (28), (32), and (33).

The Moments of the Distributions

The generalized characteristic function of (20) offers a convenient way of obtaining the moments since the nth moment is the coefficient of the term $(i\xi)^n/n!$ Table I gives the expressions for the first three moments, and the second and third moment about the mean for the three cases given above.

If the nth moment for the density function $W'(\beta)$ is desired, it can be obtained by multiplying the nth moment for $W(\phi)$ by σ^{2n}.

[5] W. Magnus and F. Oberhettinger, "Special Functions of Mathematical Physics," Chelsea Publishing Co., New York, N. Y., p. 118; 1949.

[6] J. S. Wishart and M. S. Bartlett, "The distribution of second order moment statistics in a normal system," *Proc. Cambridge Phil. Soc.*, vol. 28, pp. 455–459; 1932.

[7] A. Erdelyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, "Table of Integral Transforms," McGraw-Hill Book Co., Inc., New York, N. Y., vol. 1, p. 119; 1954.

[8] A. Erdelyi, W. Magnus, F. Oberhettinger, and F. G. Tricomi, "Higher Transcendental Functions," McGraw-Hill Book Co., Inc., New York, N. Y., vol. 1, p. 264; 1953.

[9] J. I. Marcum, "A Statistical Theory of Target Detection by Pulsed Radar," RAND Corp., Santa Monica, Calif., Res. Memo. RM 753, Math. Appendix, pp. 39–46; 1952. *Note added in proof:* More detailed discussions of series expansions of the frequency for the special case $k = 1$ are given by C. C. Craig, "On the frequency function of xy," *Ann. Math. Statistics*, vol. 7, pp. 1–15; March, 1936, and L. A. Aroian, "The probability function of the product of two normally distributed variables," *Ann. Math. Statistics*, vol. 18, pp. 265–271; June, 1947.

TABLE I
MOMENTS FOR $W(\phi)$

	Case $n - n$	Case $n - s$	Case $s - s$
Mean	$k\rho$	0	R
Second Moment	$k(1 + \rho^2) + k^2\rho^2$	$k + R$	$k + 2R + R^2$
Third Moment	$\rho k(k + 2)[3 + \rho^2(k + 1)]$	0	$R[(6 + 3k) + 6R + R^2]$
Variance	$k(1 + \rho^2)$	$k + R$	$k + 2R$
Third Moment About the Mean	$2k\rho(3 + \rho^2)$	0	$6R$

The Saddle Point Approximation

By using the definitions (12), (13), (24), and (25), 22) may be written

$$W(\phi) = \frac{L}{2\pi}\int_{-\infty}^{\infty} \exp\{-kF(z)\}\, dz \tag{34}$$

here

$$F(z) = \tfrac{1}{2}\log(1 + z^2) + iz\eta + \gamma\frac{hz^2 - icz}{1 + z^2}. \tag{35}$$

Along the imaginary axis, $F(z)$ has poles at $z = \pm i$, nd has a single stationary point at some imaginary value f z between $+i$ and $-i$. The saddle point approximation s obtained by shifting the path of integration until it asses through this stationary point and then taking a 'aylor's series expansion of F about the stationary point. Ve define Q as the real root (with $Q^2 \leqq 1$) of $F'(-iQ) = 0$, where the prime denotes differentiation with respect to z. Ve also let

$$E = F(-iQ), \tag{36}$$

$$D = 2F''(-iQ), \tag{37}$$

nd substitute $z = -iQ + 2u/\sqrt{kD}$. Then (34) becomes

$$W(\phi) \cong \frac{L}{\pi\sqrt{kD}}\int_{-\infty}^{\infty} e^{-kE-u^2} \cdot \exp\left\{-\frac{4}{3}\frac{u^3F'''(-iQ)}{D\sqrt{kD}} - \frac{2}{3}\frac{u^4F^{IV}(-iQ)}{kD^2} + \cdots\right\} du. \tag{38}$$

Since the integrand is very small except when u is small, he terms which involve the third and higher derivatives nay be treated as small correction terms. The integration over u is carried out by first expanding the second exponential in (38) as a power series. The result of this method of approximation is most conveniently expressed by defining both $W(\phi)$ and ϕ in terms of the common parameter Q. The required set of equations is

$$W(\phi) \cong \frac{L}{\sqrt{\pi kD}} e^{-kE} \cdot\left\{1 + \frac{1}{12k}(9HG^{-2} - 10\,NG^{-3}) + \cdots\right\}, \tag{39}$$

$$\frac{\phi}{k(1 - \rho^2)} = \eta = \left\{\frac{Q}{1 - Q^2} + \frac{\gamma}{(1 - Q^2)^2}[2hQ + c(1 + Q^2)]\right\} \tag{40}$$

where

$$E = \tfrac{1}{2}\ln(1 - Q^2) + \frac{Q^2}{1 - Q^2} + \frac{\gamma[hQ^2(1 + Q^2) + 2cQ^3]}{(1 - Q^2)^2}, \tag{41}$$

$$G = (1 - Q^4) + 2\gamma h(1 + 3Q^2) + 2\gamma c(3Q + Q^3), \tag{42}$$

$$D = 2G(1 - Q^2)^{-3}, \tag{43}$$

$$H = (1 - Q^2)^2(1 + 6Q^2 + Q^4) + 4\gamma(1 - Q^2) \cdot(h + 5cQ + 10hQ^2 + 10cQ^3 + 5hQ^4 + cQ^5), \tag{44}$$

$$N = (1 - Q^2)\{(3Q + Q^3)(1 - Q^2) + 3\gamma(c + 4hQ + 6cQ^2 + 4hQ^3 + cQ^4)\}^2. \tag{45}$$

Since (40) is a quartic in Q, standard methods could be used to write Q as an explicit function of ϕ and γ, but the numerical work is much simpler if the parametric arrangement is used. The density functions for the three cases of primary interest may be computed from the above formulas with the following choices for the parameters:

$$W_{nn}(\phi) : h = c = 0,$$

$$W_{ns}(\phi) : h = \tfrac{1}{2}, \quad c = 0, \quad \rho = 0,$$

$$W_{ss}(\phi) : h = c = 1, \quad \rho = 0.$$

Accuracy and Limiting Forms

Fig. 2 shows the values of $W_{ss}(\phi)$, $W_{sn}(\phi)$, and $W_{nn}(\phi)$ as computed from (39) and (40) for the choice $k = 11$, $\rho = 0$, $R = 11$. These numerical results were checked against the exact expression (28) for $W_{nn}(\phi)$ over the entire range, and against the series (30) for $W_{sn}(\phi)$ at two points. In all cases the error was less than 1 per cent. For large values of R, the convergence rate of the series (32) and (33) is slow enough to make machine computation of the series almost a necessity, so no comparison checks on the values of $W_{ss}(\phi)$ have been made.

The accuracy of (39) can be increased, if desired, by carrying the Taylor series expansion in (38) out to higher

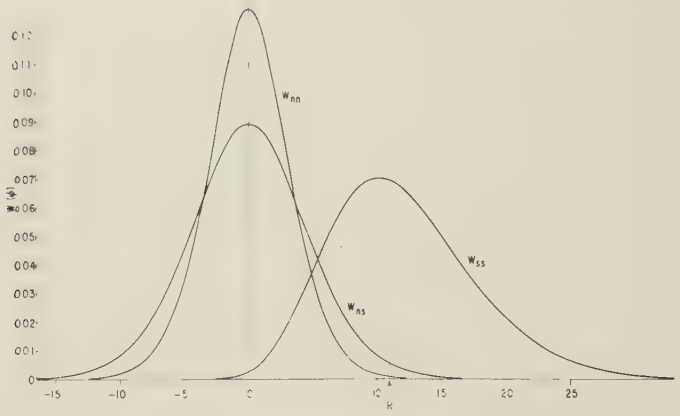


Fig. 2—Probability density function $W(\phi)$.

derivatives. However, the accuracy of (39) as it stands is sufficient for most practical purposes.

By the Central Limit Theorem, the above density functions should approach a Gaussian form for k large. In the following discussion, $\rho = 0$ for all three cases. The Gaussian forms obtained by treating Q as small in (39)–(45) are

$$W_{nn}(\phi) \rightarrow \frac{1}{\sqrt{2\pi k}} e^{-\phi^2/2k}, \tag{46}$$

$$W_{sn}(\phi) \rightarrow \frac{1}{\sqrt{2\pi(k+R)}} e^{-\phi^2/2(k+R)}, \tag{47}$$

$$W_{ss}(\phi) \rightarrow \frac{1}{\sqrt{2\pi(k+2R)}} e^{-(\phi-R)^2/2(k+2R)}. \tag{48}$$

The Gaussian form is accurate only near the peak of the density function. The nature of the initial deviation away from the pure Gaussian form is easily found by an expansion of (40)–(45) in powers of Q. For example, the Gaussian approximation for $W_{ss}(\phi)$, above, should be multiplied by a factor

$$\left\{1 + \frac{R(\phi - R)^3}{(k+2R)^3} - \frac{3R(\phi - R)}{(k+2R)^2} + \cdots\right\}.$$

These correction terms form the beginning of an Edgeworth series,[10] and further terms in the series could be constructed from the known moments. Once the distribution starts to depart from the Gaussian form, the deviation can become quite large, and as will be seen below, the shape of the tails is more nearly exponential than Gaussian. A series of the Edgeworth type does not present an easily interpretable picture of the shape of the tails of the distribution, because the value of k required to keep the fractional error within some specified limit increases with ϕ.

This difficulty does not appear in our (39). Note that the quantities H/G^2 and N/G^3 are bounded for all Q in the range $-1 \leqq Q \leqq 1$, so that the first order correction term in (39) is of order $1/k$ even for ϕ infinite. The shape of the tails can therefore be determined by letting Q^2 approach unity in (39)–(45). Neglecting terms of order $1/k$ compared to unity, the three distributions of special interest become, for $\phi \gg \sqrt{k}$,

$$W_{nn}(\phi) \cong \frac{1}{\sqrt{4\pi k}} (\phi/k)^{k/2-1} e^{-\phi+k/2}, \tag{49}$$

$$W_{sn}(\phi) \cong \frac{1}{\sqrt{8\pi R}} \left(\frac{\phi}{R}\right)^{(k-3)/4} e^{-\phi+\sqrt{\phi R}-3/8R}, \tag{50}$$

$$W_{ss}(\phi) \cong \frac{1}{\sqrt{32\pi R}} \left(\frac{\phi}{4R}\right)^{(k-3)/4} e^{-\phi+2\sqrt{\phi R}-R+k/2}, \tag{51}$$

$$W_{ss}(-\phi) \cong \frac{1}{\sqrt{32\pi R}} \left(\frac{4\phi + R + k}{4k}\right)^{k-3/2} \cdot e^{-\phi-R/2+k/2}. \tag{52}$$

The contrast between these limiting forms, valid in the tails of the distributions, and the limiting forms (46)–(48), which hold near the peak of the distribution, is quite sharp. The transition between the two limiting forms is easy to illustrate in the case of $W_{nn}(\phi)$, for in that case $\gamma = 0$ and (40) becomes a quadratic which is easily solved for Q as a function of η. The variation in $W_{nn}(\phi)$ depends, for large k, almost entirely on the dominating exponential term e^{-kE}, and for this case E becomes

$$E_{nn} = \left(\frac{\sqrt{1+4\eta^2}-1}{2}\right) + \tfrac{1}{2} \log \left(\frac{\sqrt{1+4\eta^2}-1}{2\eta^2}\right). \tag{53}$$

For η small, $E_{nn} \rightarrow \frac{1}{2}\eta^2$, corresponding to the Gaussian shape in (46), while for η large, $E_{nn} \rightarrow \eta - \frac{1}{2} - \frac{1}{2}\log \eta$, in agreement with (49). For the cases *ss* and *sn* it is not possible to write as simple an explicit expression for E, but it is clear from (40) and (41) that E depends on k only through the ratios η and γ.

Conclusions

In this article we have derived several alternate expressions for the density function for correlators that have their reference signals corrupted by additive noise. The series expansions for the density functions are rather slowly converging, and except for some special cases are not very convenient for numerical computations. The asymptotic expressions which result from a saddle point integration are best written by defining both the density function $W(\phi)$ and the variable ϕ as functions of a parametric variable Q. When this parametric form is used our results are accurate to terms of order $1/k$ over the entire range of ϕ. Explicit expressions developed from the parametric formulation are less accurate, but show a transition from the expected Gaussian shape near the peak of the density function to a more nearly exponential shape in the tails of the distribution.

[10] H. Cramér, "Mathematical Methods of Statistics," Princeton University Press, Princeton, N. J., sect. 17.7; 1951.

Demodulation of a Phase-Modulated Noise Carrier*

PHILLIP BELLO†, ASSOCIATE MEMBER, IRE

Summary—**In this paper, an analysis is made of a communication system in which the information-bearing signal phase modulates a Gaussian noise carrier. The effect of additive Gaussian noise and linear filtering on the first-order statistics of the receiver output noise and on the character of the output signal are determined. It is shown that with regard to determining the distortion of the output signal, the system may be replaced by a single linear filter whose input is the modulated signal impressed on a sinusoidal rather than noise carrier. In this way, conventional FM techniques may be used for the determination of signal distortion.**

I. Introduction

THE conventional information-carrying type of signal is the modulated periodic time function. Recently, however, consideration has been given to the use of noise and noise-like signals as carriers of information.[1-3] Moreover, in conventional fading communication channels, a phase-modulated noise carrier is actually received even if a PM *sine wave* carrier is transmitted. To counteract the effect of the fading mechanism, a pilot tone may be transmitted. Due to the fading medium, the pilot tone is received as a narrow-band noise process with an amplitude and PM the same (assuming nonselective fading) as that impressed upon the PM sine wave carrier. By mixing this received pilot tone with the received PM carrier, one may presumably remove the noisy PM due to the fading mechanism.

It is the purpose of this paper to investigate the first-order statistics of the noise appearing at the output of the demodulator for a PM signal employing a noise carrier. Specifically, it is desired to determine the effects of linear distortion and additive noise on the character of noise and signal at the demodulator output. Since the carrier is narrow-band noise, the reference for the phase detector must be as near as possible an unmodulated replica of the noise carrier. The essentials of the system to be analyzed are shown in Fig. 1. As shown, the noise carrier is phase modulated by a signal $\varphi(t)$ and then distorted by passage through linear filter 2. The resulting waveform has Gaussian noise $n_2(t)$ added to it (possibly receiver noise) and is then presented to the demodulator. The reference signal for the phase detector is shown as a replica of the original noise carrier, which has been distorted by passage through linear filter 1 and perturbed by additive noise $n_1(t)$. On the assumption that the noise carrier is a stationary Gaussian process, and $\varphi(t)$ is deterministic, one may show that the process at the output of filter 2 is also Gaussian, although nonstationary. It then becomes clear that the demodulated output (assuming an ideal phase detector) is just the phase difference between a stationary and a correlated nonstationary narrow-band Gaussian process. Thus, the ensemble statistical properties of the phase difference process will be a function of time.

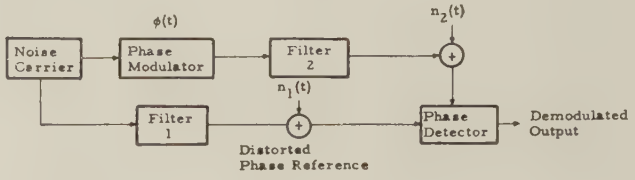


Fig. 1—Functional block diagram of system to be analyzed.

In the following section, the probability density function of the phase detector output will be derived. It may readily be seen that if filters 1 and 2 are identical and have pass bands much wider than the bandwidths of their input signals, then, in the absence of additive noise, the output of an ideal phase detector will be just $\varphi(t)$, the input PM. If $\varphi(t)$ is zero and the additive noise is absent, the phase detector output will contain noise if there is any dissimilarity (apart from a gain change) in filters 1 and 2. Moreover, as will be shown subsequently, even if filters 1 and 2 are identical and the additive noise is absent, phase detector output will contain noise if $d\varphi(t)/dt \neq 0$. Such noise may properly be called "self" noise since it is present under "no interference" conditions and is due, moreover, to the noisy nature of the carrier. It will be presumed that the phase detector is conventional, in the sense that it is a periodic function of its input (ξ) with a period of 2π, *i.e.*, that the phase detector cannot distinguish between $\xi = \delta$ and $\xi = \delta + 2k\pi$, where $|\delta| < \pi$ and $k = 0, \pm 1, \pm 2$, etc. The periodic character of the phase detector characteristic may be accounted for automatically in the calculation of first-order output statistics if in the calculation of the density function of $\xi(t)$, all values of $\xi(t)$ outside the interval $(-\pi, \pi)$ are referred to this interval modulo 2π. When this is done, all values of the phase detector input are confined to the interval $(-\pi, \pi)$ and it is only necessary to examine the phase detector characteristic within this interval.

In order for the PM system of Fig. 1 to be usable, it is necessary that the noise output of the phase detector be small compared with the signal output. It will be assumed that the PM is confined to the interval $-\pi + G \leq \varphi \leq \pi - G$, where $G > 0$ is a guard band inserted

* Received by the PGIT, April 15, 1960; revised manuscript received, August 26, 1960.

† Appl. Res. Lab., Sylvania Electronic Systems, a division of Sylvania Electric Products, Inc., Waltham, Mass.

[1] B. M. Horton, "Noise modulated distance measuring system," Proc. IRE, vol. 47, pp. 821–828; May, 1959.

[2] A. A. Kharkevich, "The transmission of signals by modulated noise," in "Telecommunications," Pergamon Press, Inc., London, England, pp. 43–47; 1957.

[3] R. Price and P. E. Green, "A communication technique for multipath channels," Proc. IRE, vol. 46, pp. 555–569; March, 1958.

because conventional phase detectors are not always linear over a full range of 2π. The presence of noise and signal distortion causes the phase difference process $\xi(t)$ to have values beyond the interval $(-\pi + G, \pi - G)$. However, presuming the modulo 2π representation for $\xi(t)$ has been used, it will still be confined to the interval $(-\pi, \pi)$.

The definition of output signal deserves some separate consideration, since the over-all PM system from input modulation to phase detector output is both nonlinear and time varying. The output signal $\varphi_0(t)$ will be defined as the ensemble average of the phase detector output, *i.e.*,

$$\varphi_0(t) \equiv \overline{F(\xi(t))}, \tag{1}$$

where $F(\xi)$ is the phase detector characteristic. It will be presumed that the phase detector characteristic is linear within the interval $(-\pi + K \leq \xi \leq \pi - K)$, *i.e.*,

$$F(\xi) = \xi \quad \text{for} \quad -\pi + K \leq \xi \leq \pi - K, \tag{2}$$

where K is a positive number less than G. This means that when filters 1 and 2 are identical with passbands much wider than the bandwidths of their input signals and when additive noise is absent, the output of the phase detector will be equal to $\varphi(t)$, just as for an ideal phase detector.

It will be subsequently demonstrated that (1) and (2) have the useful property of producing an output signal which is independent of the additive noises in the direct and reference channels.

A most useful property of (1) and (2) is the result that $\varphi_0(t)$ may be computed as the phase response of a linear filter (called the equivalent linear filter) to the input PM $\varphi(t)$ on a *sinusoidal* rather than a *noise* carrier.

To simplify subsequent analysis, the complex representation of real waveforms will be used.[4-7] Specifically, the concept of a "complex" envelope for a narrow-band time function will be used. A real process $S(t)$ has a representation in the form

$$S(t) = \text{Re}\ \{e(t)e^{j\omega_0 t}\},$$

where $e(t)e^{j\omega_0 t}$ has a spectrum confined to positive frequencies (Re $\{\ \}$ is the usual real part notation). When $S(t)$ is narrow-banded, $e(t)$ will be called the "complex" envelope of $S(t)$, since then the magnitude of $e(t)$ may be identified with the conventional envelope of $S(t)$, while the angle of $e(t)$ may be identified with the conventional phase variation of $S(t)$ about the carrier phase $\omega_0 t$.

In the discussion that follows, it will be possible to deal entirely with complex envelopes without reference to absolute center frequencies.

II. Probability Density Function of Phase Detector Output

In this section, there will be derived the probability density function (PDF) for the output of the phase detector of Fig. 1. To facilitate the exposition, Fig. 1 is redrawn as Fig. 2 with various pertinent complex envelopes indicated.

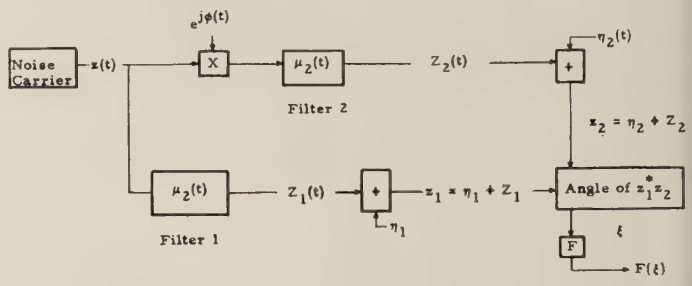


Fig. 2—Pertinent complex envelopes in system to be analyzed.

The complex envelope of the noise carrier is $z(t)$. Note that the operation of PM of the noise carrier by $\phi(t)$ is represented in terms of complex envelopes as a multiplication by $e^{j\phi(t)}$. The complex envelopes of the impulse responses of filters 1 and 2 are given by $\mu_1(t)$ and $\mu_2(t)$, respectively. It is readily demonstrated that in the case of a narrow-band filter with a narrow-band input (as assumed here) the output complex envelope is one-half the convolution of the input complex envelope with the filter impulse response complex envelope. Thus the complex envelope of filter 1 output $Z_1(t)$ is given by

$$Z_1(t) = \tfrac{1}{2}\mu_1(t) \otimes z(t), \tag{3}$$

while the complex envelope of filter 2 output $Z_2(t)$ is given by

$$Z_2(t) = \tfrac{1}{2}\mu_2(t) \otimes z(t)e^{j\phi(t)}, \tag{4}$$

where the symbol $\otimes$ denotes convolution. The phase detector reference complex envelope $z_1(t)$ is given by

$$z_1(t) = \eta_1(t) + Z_1(t), \tag{5}$$

where $\eta_1(t)$ is the complex envelope of an additive noise. Similarly, the phase detector input $z_2(t)$ is given by

$$z_2(t) = \eta_2(t) + Z_2(t), \tag{6}$$

where $\eta_2(t)$ is the complex envelope of another independent additive noise. The phase difference process $\xi(t)$ is readily seen to be expressible as the angle of $z_1^* z_2$. Thus, the phase detector output is shown in Fig. 2 to be computed by a cascade of two devices, one of which extracts the angle of $z_1^* z_2$ and feeds it to the other, a no-memory nonlinear device with characteristic $F(\xi)$.

Presuming that the noise carrier is an ergodic Gaussian process, the $z(t)$ is a complex-valued Gaussian ergodic process.[8] If it is assumed that the modulation is a deter-

[4] J. Dugundji, "Envelopes and pre-envelopes of real waveforms," IRE Trans. on Information Theory, vol. IT-5, pp. 53–57; March, 1958.

[5] R. Arens, "Complex processes for envelopes of normal noise," IRE Trans. on Information Theory, vol. IT-4, pp. 204–207; September, 1957.

[6] P. M. Woodward, "Probability and Information Theory," McGraw-Hill Book Co., Inc., New York, N. Y.; 1953.

[7] D. Gabor, "Theory of communications," *J. IEE*, vol. 93, pt. III, pp. 429–457; 1946.

[8] For a definition of the complex valued Gaussian process and some of its properties, see J. L. Doob, Stochastic Processes," John Wiley and Sons, Inc., New York, N. Y., p. 71; 1953.

inistic function of time, then $z(t)e^{j\phi(t)}$ is still normally istributed, although nonstationary. Since any linear eterministic operation on a normal process leaves the ormal character unchanged, it is clear that $Z_1(t)$ and $_2(t)$ [and thus $z_1(t)$ and $z_2(t)$] are normal complex rocesses, although $Z_2(t)$ [and $z_2(t)$] is nonstationary.

It will be convenient to denote ξ at a given time instant s ξ_t. The PDF of ξ_t can be expressed as

$$W(\xi_t) = \int_0^\infty \int_0^\infty \int_0^{2\pi} \int_0^{2\pi} \delta(\xi_t - [\theta_2 - \theta_1]) \cdot F(r_1 e^{j\theta_1}, r_2 e^{j\theta_2}) r_1 r_2 \, dr_1 \, dr_2 \, d\theta_1 \, d\theta_2, \quad (7)$$

here

$$z_1 = r_1 e^{j\theta_1} \qquad z_2 = r_2 e^{j\theta_2} \quad (8)$$

nd $\delta(\xi - A)$ is a unit impulse at $\xi = A$. The function $'(z_1, z_2)$ is the joint PDF for the complex normal variates $, z_2$. Assuming that z_1, z_2 have zero mean, this function s given by

$$F(z_1, z_2) = \frac{\exp\left[-\frac{1}{(1-\rho_t^2)}\left(\frac{|z_1|^2}{\overline{|z_1|^2}} + \frac{|z_2|^2}{\overline{|z_2|^2}} - \frac{\mathrm{Re}\{\gamma_t z_1 z_2^*\}}{\sqrt{\overline{|z_1|^2}\,\overline{|z_2|^2}}}\right)\right]}{(2\pi)^2 \overline{|z_1|^2}^{\frac{1}{2}} \overline{|z_2|^2}^{\frac{1}{2}} (1-\rho_t^2)}, \quad (9)$$

here $\overline{|z_k|^2}$ is the ensemble average of $z_k z_k^*$, γ_t is the nor- nalized complex cross-correlation coefficient,

$$\gamma_t = \frac{\overline{z_1^* z_2}}{\sqrt{\overline{|z_1|^2}\,\overline{|z_2|^2}}} = \rho_t e^{j\beta_t}, \quad (10)$$

nd ρ_t, β_t are its magnitude and angle.

The integrations with respect to θ_1 and θ_2 in (7) are rivial. The integrations with respect to r_1 and r_2 may e carried out with the aid of the integral,

$$\int_0^\infty \int_0^\infty XY \exp\left[-(X^2 + Y^2 + 2XY\cos\epsilon)\right] dX\, dY = \tfrac{1}{4}\csc^2\epsilon(1 - \epsilon\cos\epsilon), \quad (11)$$

which may be found in Rice.[9] Considering that ξ_t is onfined to the interval $-\pi < \xi_t < \pi$ (in the manner ndicated in the Introduction), it is readily determined hat

$$W(\xi_t) = \frac{1}{2\pi}\left[\frac{1-\rho_t^2}{1-\rho_t^2\cos^2(\xi_t - \beta_t)}\right] \cdot\left[1 + \frac{\rho_t\cos(\xi_t - \beta_t)[\pi - \cos^{-1}\{\rho_t\cos(\xi_t - \beta_t)\}]}{\sqrt{1-\rho_t^2\cos^2(\xi_t - \beta_t)}}\right] \quad (12)$$

$$0 < \cos^{-1}[\rho_t\cos(\xi_t - \beta_t)] < \pi$$

$$-\pi < \xi_t < \pi.$$

[9] S. O. Rice, "Mathematical analysis of random noise," in Noise and Stochastic Processes," N. Wax, Ed., Dover Publications, nc., New York, N. Y., p. 207; 1954.

It is important to note that the PDF of ξ_t is entirely fixed by $\gamma_t = \rho_t \exp[j\beta_t]$. When $\rho_t = 0$, the PDF of ξ_t becomes uniformly distributed. As ρ_t increases from zero, the PDF begins to peak at $\xi_t = \beta_t$. As ρ_t approaches 1 (it can never be greater than 1), $W(\xi_t)$ becomes an impulsive type PDF centered on $\xi_t = \beta_t$. In the limit $\rho_t = 1$, $W(\xi_t)$ becomes a unit impulse located at $\xi_t = \beta_t$. Fig. 3 shows a plot of $W(\xi_t)$ for $\rho_t = 0.95$, 0.995 and 0.999.

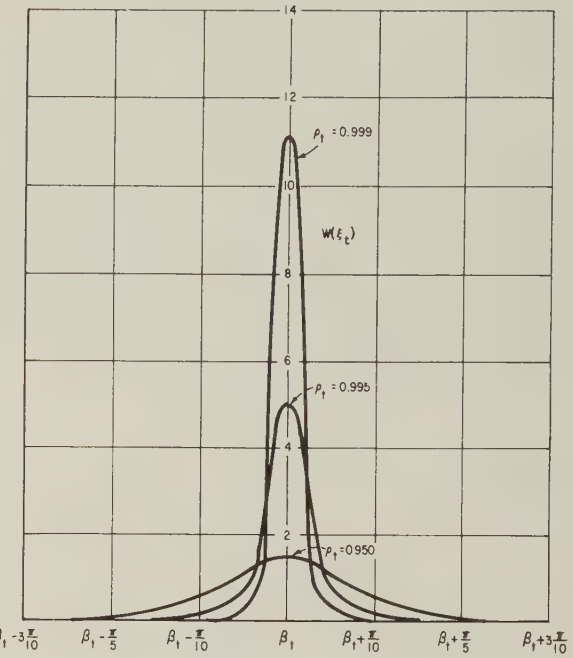


Fig. 3—Density function for ξ_t; $\rho_t = 0.95$, 0.995, 0.999.

Examination of Fig. 3 shows that, as evidenced by the spread of the density function, a non-negligible phase noise is present in the phase difference process even for $\rho_t = 0.950$. For $\rho_t \geq 0.950$, the approximation

$$W(\xi_t) \approx \frac{\alpha_t}{2[1 + \alpha_t^2(\xi_t - \beta_t)^2]^{3/2}} \quad ; -\pi < \xi_t < \pi \quad ; \alpha_t \gg 1, \quad (13)$$

where

$$\alpha_t = \frac{\rho_t}{\sqrt{1-\rho_t^2}} = \tan\{\sin^{-1}\rho_t\}, \quad (14)$$

may be used. In fact, the approximation represented in (13) would be graphically indistinguishable from the corresponding curves of Fig. 3 if plotted for the same values of ρ_t. However, there is one situation in which (13) will not give an adequate representation to $W(\xi_t)$, even for $\rho_t \geq 0.950$. This occurs when β_t is close enough to π or $-\pi$. To be more precise, if L denotes the length of the interval over which (13) has significant nonzero values (according to some reasonable criterion), then this approximation should only be used if $|\beta_t| < \pi - L/2$. For values of ρ_t near 1, L will become small enough so that the condition $|\beta_t| < \pi - L/2$ will not be too restrictive. It will be reasoned now that, in fact, $|\beta_t|$

cannot become appreciably greater than $\pi - G$ when

$$\frac{L}{2} < G - K \tag{15}$$

and the output signal distortion is small.

The reasoning is as follows. If (15) is satisfied, then the density function $W(\xi_t)$ is confined to the linear portion of the phase detector characteristic and is closely given by the unimodal function of (13). Therefore,

$$\phi_0 \equiv \overline{F(\xi)} = \bar{\xi} \approx \beta_t. \tag{16}$$

If signal distortion is small, then

$$| \beta_t | \approx | \phi_0 | \approx | \phi(t) | < \pi - G, \tag{17}$$

as was to be demonstrated.

For a fixed pair G, K, (15) will be satisfied automatically if ρ_t is sufficiently close to 1 or, equivalently, if $\alpha_t \gg 1$. It will be assumed in the remainder of this paper that (13) yields a satisfactory approximation to $W(\xi_t)$. Without going into the algebra, it is readily demonstrated that if the output noise N_t is defined as

$$N_t = \xi_t - \phi_0(t), \tag{18}$$

then

$$\overline{| N_t |} \approx \frac{1}{\alpha_t},$$

$$\overline{N_t^2} \approx \frac{\log 2\pi\alpha_t - 1}{\alpha_t^2}, \tag{19}$$

$$\Pr(| N_t | \leq \phi) = \frac{\alpha_t \phi}{\sqrt{1 + \alpha_t^2 \phi^2}},$$

provided $\alpha_t(\pi - | \beta_t |) \gg 1$.

Since the average magnitude of the fluctuations of ξ_t about its average value ϕ_0 is just $1/\alpha_t$, and since the peak value of $W(\xi_t)$ [see (13)] is $\alpha_t/2$, it is convenient to visualize the density function of ξ_t as being contained within a rectangle $2/\alpha_t$ wide and $\alpha_t/2$ high centered on ϕ_0. Actually, the area of $W(\xi)$ contained within the limits $\phi_0 - 1/\alpha_t$ and $\phi_0 + 1/\alpha_t$ is 0.707, as is readily seen by use of the last part of (19).

The expression for the mean squared value of N_t is not as simple as that for the mean absolute value of N_t. Because of the simplicity of the latter expression, it has been found more convenient for use as a measure of the spread of ξ_t (or N_t). Thus, if one assumes a peak output signal of ϕ_m radians, a simple measure of the output peak signal-to-noise ratio is $\phi_m / \overline{| N_t |} = \alpha_t \phi_m$. On this basis, the curves shown in Fig. 3 for $\rho_t = 0.950$, 0.995, and 0.999 correspond to signal-to-noise ratios of 3, 10, and 22, respectively, for a peak output signal of 1 radian.

To summarize briefly: if it is assumed that

$$\alpha_t(\pi - | \beta_t |) \gg 1,$$

then the determination of the output signal involves an evaluation of β_t, the angle of

$$\gamma_t = \frac{\overline{z_1^* z_2}}{\sqrt{\overline{| z_1 |^2}\ \overline{| z_2 |^2}}},$$

(the normalized complex cross-correlation coefficient between the PM signal and the phase reference signal) while the determination of the output noise level (on an ensemble basis) requires an evaluation of $\rho_t = | \gamma_t |$.

III. Evaluation of the Complex Correlation Coefficient

In this section, an expression for γ_t will be derived, showing its dependence upon the input modulation, the autocorrelation function of the noise carrier, and the impulse responses of the filters.

If it is assumed that η_1 and η_2 are independent of one another and of z_1 and z_2, then

$$\begin{aligned} \overline{z_1^* z_2} &= \overline{Z_1^* Z_2} \\ \overline{z_1^* z_1} &= \overline{| Z_1 |^2} + \overline{| \eta_1 |^2} = 2P_{S_1} + 2P_{n_1} \\ \overline{z_2^* z_2} &= \overline{| Z_2 |^2} + \overline{| \eta_2 |^2} = 2P_{S_2} + 2P_{n_2}, \end{aligned} \tag{20}$$

where the ensemble powers $P_{S_1}, P_{S_2}, P_{n_1}, P_{n_2}$ are as defined below:

$$\begin{aligned} P_{S_1} &= \tfrac{1}{2}\overline{| Z_1 |^2}; \qquad P_{S_2} = \tfrac{1}{2}\overline{| Z_2 |^2}; \\ P_{n_1} &= \tfrac{1}{2}\overline{| \eta_1 |^2}; \qquad P_{n_2} = \tfrac{1}{2}\overline{| \eta_2 |^2}. \end{aligned} \tag{21}$$

It follows that γ_t may be written as

$$\gamma_t = \hat{\gamma}_t\, d_t, \tag{22}$$

where

$$\hat{\gamma}_t = \frac{\overline{Z_1^* Z_2}}{\sqrt{\overline{| Z_1 |^2}\ \overline{| Z_2 |^2}}}\, d_t = \frac{1}{\sqrt{\left(1 + \frac{P_{n_1}}{P_{S_1}}\right)\left(1 + \frac{P_{n_2}}{P_{S_2}}\right)}}. \tag{23}$$

Thus, γ_t is representable as the product of the complex correlation coefficient with no additive noise, $\hat{\gamma}_t$, by a degradation factor $d_t < 1$. This degradation factor depends only upon the ensemble signal-power-to-noise-power ratios at the two inputs to the phase detector. The reason for calling d_t a degradation factor is clear from our discussion of the probability density function of ξ_t. Any effect which decreases the magnitude of γ_t is a degrading effect, since the output noise is thereby increased. We will turn our attention almost exclusively to an evaluation of γ_t, since d_t is only a function of signal-to-noise ratios before phase detection and may be dealt with afterward. Note that since d_t is a real positive quantity,

$$\varphi_0(t) = \text{angle of } \gamma_t = \text{angle of } \hat{\gamma}_t. \tag{24}$$

Thus, the output signal $\varphi_0(t)$ [*as defined in* (1)] *is not dependent upon the additive noise and may be calculated directly from* $\hat{\gamma}_t$.

Turning now to the evaluation of $\hat{\gamma}_t$, we note from Fig. 2 that

$$Z_1(t) = \tfrac{1}{2}z(t) \otimes \mu_1(t) = \tfrac{1}{2}\int_0^\infty \mu_1(\sigma) z(t - \sigma)\, d\sigma \tag{25}$$

$$Z_2(t) = \tfrac{1}{2}z(t)e^{j\varphi(t)} \otimes \mu_2(t) = \tfrac{1}{2}\int_0^\infty \mu_2(\beta) z(t - \beta) e^{j\varphi(t-\beta)}\, d\beta.$$

Consequently,

$$\overline{Z_1^* Z_2} = \tfrac{1}{4}\int_0^\infty \int_0^\infty \mu_1^*(\sigma)\mu_2(\beta)e^{j\varphi(t-\beta)} \cdot \overline{z^*(t-\sigma)z(t-\beta)}\, d\sigma\, d\beta$$

$$\overline{Z_1^* Z_1} = \tfrac{1}{4}\int_0^\infty \int_0^\infty \mu_1^*(\sigma)\mu_1(\beta)\, \overline{z^*(t-\sigma)z(t-\beta)}\, d\sigma\, d\beta \quad (26)$$

$$\overline{Z_2^* Z_2} = \tfrac{1}{4}\int_0^\infty \int_0^\infty \mu_2^*(\sigma)\mu_2(\beta)e^{j[\varphi(t-\beta)-\varphi(t-\sigma)]} \cdot \overline{z^*(t-\sigma)z(t-\beta)}\, d\sigma\, d\beta$$

(assuming the validity of the interchange of integration and ensemble averaging).

The statistical averages inside these integrals require a little discussion.

It will be recalled that the $z(t)$ process is stationary. The ensemble autocorrelation function $\overline{z^*(t_1)z(t_1+\tau)}$ is then a function of τ only, *i.e.*,

$$\overline{z^*(t_1)z(t_1+\tau)} = R(\tau). \quad (27)$$

In terms of the real and imaginary parts of $z(t)$, $R(\tau)$ may be expressed as

$$R(\tau) = 2R_{xx}(\tau) + j2R_{xy}(\tau), \quad (28)$$

where the even function $R_{xx}(\tau)$ is the common autocorrelation function of $x = \text{Re}\,\{z\}$ and $y = \text{Im}\,\{z\}$ and the odd function $R_{xy}(\tau)$ is the crosscorrelation function between x and y.

Using the autocorrelation function $R(\tau)$, we may write the averages in (26) as

$$\overline{Z_1^* Z_2} = \tfrac{1}{4}\int_0^\infty \int_0^\infty \mu_1^*(\sigma)\mu_2(\beta)e^{j\varphi(t-\beta)}R(\sigma-\beta)\, d\sigma\, d\beta,$$

$$\overline{Z_1^* Z_2} = \tfrac{1}{4}\int_0^\infty \int_0^\infty \mu_1^*(\sigma)\mu_1(\beta)R(\sigma-\beta)\, d\sigma\, d\beta = 2P_{S_1}, \quad (29)$$

$$\overline{Z_2^* Z_2} = \tfrac{1}{4}\int_0^\infty \int_0^\infty \mu_2^*(\sigma)\mu_2(\beta)e^{j[\varphi(t-\beta)-\varphi(t-\sigma)]} \cdot R(\sigma-\beta)\, d\sigma\, d\beta = 2P_{S_2},$$

and obtain the following general expression for $\hat{\gamma}_t$:

$$\hat{\gamma}_t = \frac{\int_0^\infty \int_0^\infty \mu_1^*(\sigma)\mu_2(\beta)e^{j\varphi(t-\beta)}R(\sigma-\beta)\, d\sigma\, d\beta}{8\sqrt{P_{S_1}P_{S_2}(t)}}. \quad (30)$$

Because of the nonstationary character of $Z_2(t)$, we note that the ensemble average power P_{S_2} is generally time varying.

It will now be demonstrated that it is possible to determine the output signal $\varphi_0(t)$ as the PM response of a linear filter to the input PM $\varphi(t)$ on a sinusoidal carrier.

To demonstrate this fact, we note from (30) that

$$\varphi_0(t) = \text{angle of} \int_0^\infty \int_0^\infty \mu_1^*(\sigma)\mu_2(\beta)e^{j\varphi(t-\beta)}R(\sigma-\beta)\, d\sigma\, d\beta. \quad (31)$$

If we define

$$\hat{\mu}_1(\beta) = \int_0^\infty \mu_1^*(\sigma)R(\sigma-\beta)\, d\sigma, \quad (32)$$

then

$$\int_0^\infty \int_0^\infty \mu_1^*(\sigma)\mu_2(\beta)e^{j\varphi(t-\beta)}R(\sigma-\beta)\, d\sigma\, d\beta = \int_0^\infty \hat{\mu}_1^*(\beta)\mu_2(\beta)e^{j\varphi(t-\beta)}\, d\beta = \hat{\mu}_1^*\mu_2 \otimes e^{j\varphi}. \quad (33)$$

Eqs. (31) and (33) show that the output signal may be regarded as the angle of the output from a linear filter with impulse response $\hat{\mu}_1^*\mu_2$ when the input is $e^{j\varphi(t)}$. The conventional real band-pass interpretation of the previous statement is that if $\cos[\omega_0 t + \varphi(t)]$ is the input to a filter with impulse response $\text{Re}\,[\hat{\mu}_1^*\mu_2 \exp(j\omega_0 t)]$, then the output signal will be $A(t)\cos[\omega_0 t + \varphi_0(t)]$, where $A(t)$ is the envelope and $\varphi_0(t)$ is the phase of the output signal. It is assumed in this latter statement that ω_0 is large enough so that the input and impulse response are narrow-band time functions. The filter with impulse response

$$h(t) = \text{Re}\,\{\hat{\mu}_1^*\mu_2 e^{j\omega_0 t}\} \quad (34)$$

will be called the equivalent linear filter. Fig. 4 depicts the determination of the output signal with the aid of the equivalent linear filter.

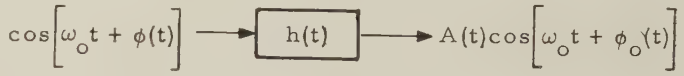


Fig. 4—Determination of output signal.

It is of interest to examine $\hat{\gamma}_t$ for three special situations:

1) no signal modulation,
2) wide-band noise carrier,
3) narrow-band noise carrier

We will consider these in order.

No Signal Modulation

This case is of special interest, since an evaluation of $\hat{\gamma}_t$ for no input modulation determines the background "self" noise of the system. When $\varphi(t)$ is zero,

$$\hat{\gamma}_t]_{\varphi=0} \equiv \hat{\gamma}_0 = \frac{\int_0^\infty \int_0^\infty \mu_1^*(\sigma)\mu_2(\beta)R(\sigma-\beta)\, d\sigma\, d\beta}{8\sqrt{P_{S_1}P_{S_2}}}. \quad (35)$$

Since $|\hat{\gamma}_0|$ does not equal one, in general, noise will appear at the phase detector output. This self noise is due to the dissimilarity in the filters, since when $\mu_1(t) = \mu_2(t)$, it is readily determined that $|\hat{\gamma}_0| = 1$.

Wide-Band Noise Carrier

If the spectrum of $z(t)$ overlaps the pass bands of filters 1 and 2 and is essentially constant in the overlap region, then, in the integrals defining $\overline{Z_1^* Z_2}$, $\overline{Z_1^* Z_1}$, and $\overline{Z_2^* Z_2}$, we may set

$$R(\tau) = K\,\delta(\tau), \quad (36)$$

where $\delta(\tau)$ is a unit impulse at $\tau = 0$, and K is the value of the spectrum of $z(t)$ at $\omega = 0$. In this case,

$$\begin{aligned}\overline{Z_1^* Z_2} &= \frac{K}{4}\int_0^\infty \mu_1^*(\sigma)\mu_2(\sigma)e^{j\varphi(t-\sigma)}\,d\sigma,\\ \overline{Z_1^* Z_2} &= \frac{K}{4}\int_0^\infty |\,\mu_1(\sigma)\,|^2\,d\sigma,\\ \overline{Z_2^* Z_2} &= \frac{K}{4}\int_0^\infty |\,\mu_2(\sigma)\,|^2\,d\sigma,\end{aligned} \tag{37}$$

and

$$\hat{\gamma}_t = \frac{\int_0^\infty \mu_1^*(\sigma)\mu_2(\sigma)e^{j\varphi(t-\sigma)}\,d\sigma}{\int_0^\infty |\,\mu_1\,|^2\,d\sigma \int_0^\infty |\,\mu_2\,|^2\,d\sigma}. \tag{38}$$

If we normalize μ_1 and μ_2 so that

$$\begin{aligned}\int_0^\infty |\,\mu_1(t)\,|^2\,dt &= 1,\\ \int_0^\infty |\,\mu_2(t)\,|^2\,dt &= 1,\end{aligned} \tag{39}$$

then

$$\hat{\gamma}_t = \int_0^\infty \mu_1^*(\sigma)\mu_2(\sigma)e^{j\varphi(t-\sigma)}\,d\sigma = \{\mu_1^*(t)\mu_2(t)\} \otimes e^{j\varphi(t)}. \tag{40}$$

Thus, we have the interesting result that in the case where the spectrum of $z(t)$ is essentially constant over the pass bands of the filters, the normalized complex correlation coefficient is just equal to the response of a linear filter with impulse response $\mu_1^*(t)\mu_2(t)$ to the excitation $e^{j\varphi(t)}$. This fact is depicted in Fig. 5. Thus, for a wide-band noise carrier, the equivalent linear filter determines both signal *and* noise output.

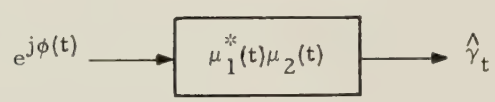


Fig. 5—Determination of $\hat{\gamma}_t$ for wide-band noise carrier.

Narrow-Band Noise Carrier

When the bandwidth of $z(t)$ is narrow compared to the bandwidth of filters 1 and 2, we may take

$$R(\tau) = 2P, \tag{41}$$

where P is the mean squared value of $z(t)$. It is then readily determined that

$$\hat{\gamma}_t = \frac{\int_0^\infty \mu_1^*(\sigma)\,d\sigma \int_0^\infty \mu_2(\beta)e^{j\varphi(t-\beta)}\,d\beta}{\left|\int_0^\infty \mu_1^*(\sigma)\,d\sigma\right| \left|\int_0^\infty \mu_2(\beta)e^{j\varphi(t-\beta)}\,d\beta\right|}. \tag{42}$$

Examination of (42) shows that $|\,\hat{\gamma}_t\,| = 1$. Consequently, there is no self noise at the phase detector output. The output signal is

$$\varphi_0(t) = \text{angle of} \int_0^\infty \mu_1^*(t)\,dt + \text{angle of} \int_0^\infty \mu_2(\beta)e^{j\varphi(t-\beta)}\,d\beta. \tag{43}$$

Thus, apart from an additive constant, the output signal is just the angle of the output of filter 2 with $e^{j\varphi(t)}$ as input.

It is interesting to note that, as one should expect, the output signal given by (44) is just what would be arrived at if a sine wave carrier were initially assumed.

IV. Determination of Self-Noise Output for a Specific Example

It is worthwhile to consider a specific case of the evaluation of $|\,\hat{\gamma}_0\,|$ in order to get a feeling for the relative effects of the noise carrier bandwidth and filter dissimilarity in producing self-noise output. If it is assumed that filters 1 and 2 are single tuned RLC filters, and that the noise carrier is obtained by passing white noise through an RLC single tuned shaping filter, then we can take

$$\begin{aligned}\mu_1(t) &= e^{-at} \quad ; \quad t < 0\\ \mu_2(t) &= e^{-(b+j\Omega)t}; \quad t < 0\\ R(\tau) &= e^{-k|\tau|}.\end{aligned} \tag{44}$$

The half-power bandwidth of filter 1 is $2a$, of filter 2 is $2b$, and of the noise carrier is $2k$ radians/second. It should be noted from (44) that in addition to differing in bandwidth by $2\,|\,b - a\,|$ radians/second, the filters differ in center frequency by Ω radians/second.

It is readily determined that

$$\begin{aligned}\overline{Z_1^* Z_2} &= \frac{1}{a+b+j\Omega}\left[\frac{1}{b+j\Omega+k} + \frac{1}{a+k}\right],\\ \overline{Z_1^* Z_1} &= \frac{1}{a(a+k)},\\ \overline{Z_2^* Z_2} &= \frac{(b+k)}{b[(b+k^2)+\Omega^2]}.\end{aligned} \tag{45}$$

After some algebraic manipulation, one finds that

$$|\,\hat{\gamma}_0\,|^2 = \left[\frac{4ab}{(b+a)^2+\Omega^2}\right] \cdot \left[\frac{1}{4}\left(\sqrt{\frac{a+k}{b+k}} + \sqrt{\frac{b+k}{a+k}}\right)^2 + \frac{\Omega^2}{4(a+k)(b+k)}\right]. \tag{46}$$

It is interesting to observe how $|\,\hat{\gamma}_0\,|^2$ varies with the carrier half-bandwidth k. In Section III, we found that for zero carrier bandwidth the "self" noise disappeared, while for large carrier bandwidth the degree of self noise approached a limiting value. One intuitively expects the degree of self noise to decrease monotonically with decreasing carrier bandwidth. This means, in our example, that $|\,\hat{\gamma}_0\,|^2$ should approach 1 asymptotically and monotonically from below, as k decreases.

Examination of (46) shows that this is indeed true. For $k = \infty$, the second term in brackets becomes unity.

hus, the first term in brackets is the $k = \infty$ value of $\gamma_0 |^2$, $| \gamma_0 |^2_\infty$. One readily determines that the second rm in brackets increases monotonically as k decreases proaching the value $1/| \gamma_0 |^2_\infty$. It is clear that a pessiistic estimate of the output noise will be obtained by suming that the carrier bandwidth overlaps the filter andwidths.

Approximate expressions valid for small self noise will ow be derived. Let the percentage difference in filter andwidths be defined as δ,

$$\delta = \frac{b - a}{a}, \tag{47}$$

nd let us choose $k = a$, *i.e.*, a carrier noise bandwidth qual to the bandwidth of filter 1. Under the assumption hat δ and Ω are small, we may obtain the following exansions for the two square bracketed terms in (46):

$$\hat{\gamma}_0 |^2_\infty = \left[\frac{4ab}{(b + a)^2 + \Omega^2}\right] = 1 - \left(\frac{\delta}{2}\right)^2 - \left(\frac{\Omega}{2a}\right)^2 + \cdots; \tag{48}$$

$$\frac{1}{4}\left(\sqrt{\frac{a + k}{b + k}} + \sqrt{\frac{b + k}{a + k}}\right)^2 + \frac{\Omega^2}{4(a + k)(b + k)}\Bigg] = 1 + \left(\frac{\delta}{4}\right)^2 + \left(\frac{\Omega}{4a}\right)^2 + \cdots.$$

hus,

$$| \hat{\gamma}_0 |^2 = 1 - \frac{3}{4}\left(\frac{\delta}{2}\right)^2 - \frac{3}{4}\left(\frac{\Omega}{2a}\right)^2. \tag{49}$$

From (14) and (19), it is clear that the square of the verage magnitude of the self noise output $(\overline{| N_t |})^2$ is iven by

$$(\overline{| N_t |})^2 = \frac{1 - | \gamma_t |^2}{| \gamma_t |^2}. \tag{50}$$

When $| \gamma_t |^2$ is near one, it is convenient to express it as

$$| \gamma_t |^2 = 1 - \epsilon_t^2, \tag{51}$$

here ϵ_t^2 is a small positive quantity. Then $(\overline{| N_t |})^2$ may e represented by the series expansion

$$(\overline{| N_t |})^2 = \epsilon_t^2[1 + \epsilon_t^2 + \epsilon_t^4 + \cdots] \approx \epsilon_t^2. \tag{52}$$

Using (49), (51), and (52), the self-noise output for he case at hand is approximately

$$(\overline{| N |})^2 = \begin{cases} \frac{3}{4}\left[\left(\frac{\delta}{2}\right)^2 + \left(\frac{\Omega}{2a}\right)^2\right] & \text{for } k = a \\ \left[\left(\frac{\delta}{2}\right)^2 + \left(\frac{\Omega}{2a}\right)\right] & \text{for } k = \infty, \end{cases} \tag{53}$$

here the subscript t has been dropped, since with $\varphi(t) = 0$ he ensemble averages are time invariant. Eq. (54) indiates that the $k = \infty$ bound is reasonably good for the ase in which the carrier bandwidth is of the order of he filter bandwidth.

If it is desired that the average magnitude of the self noise should not exceed 0.05 radians for wide-noise carrier bandwidth, then from (53) it is seen that

$$\sqrt{\left(\frac{\delta}{2}\right)^2 + \left(\frac{\Omega}{2a}\right)^2} < 0.05. \tag{54}$$

Certainly then, the filter bandwidths should not differ by more than 10 per cent and the offset in center frequencies should not be more than 5 per cent of the filter bandwidth if there is to be any possibility of the average magnitude of self noise being less than 0.05 radians. The tolerances only increase by a factor of 1.15 if one assumes $k = a$ instead of $k = \infty$.

V. Evaluation of Self Noise for General Slightly Dissimilar Channel Filtering

In this section, attention will be focused on the determination of the output self-noise level. It will be assumed that small dissimilarities exist in the filtering on the direct and reference channels. For simplicity of analysis, the noise carrier bandwidth will be assumed wide compared with the filter bandwidth. In view of the analysis in Section IV, one may expect the resulting upper bound on self-noise performance to be useful for noise-carrier bandwidths at least as small as the filter bandwidths.

Various distortions may be assumed in one filter relative to the other. Three simple distortions which are easy to characterize analytically are bandwidth, center frequency, and delay differences.

The complex correlation coefficient for large carrier bandwidth and zero signal modulation is given by

$$\hat{\gamma}_0 = \int_0^\infty \mu_1^*(t)\mu_2(t)\, dt, \tag{55}$$

on the assumption that $| \mu_1 |^2$ and $| \mu_2 |^2$ have been normalized to unit area. According to Parseval's theorem, an equivalent frequency domain expression is

$$\hat{\gamma}_0 = \int_{-\infty}^\infty U_1^*(f)U_2(f)\, df \tag{56}$$

where

$$U_k(f) = \int_0^\infty \mu_k(t)e^{-i2\pi ft}\, dt; \qquad k = 1, 2. \tag{57}$$

If filter 2 differs from filter 1 only by a small percentage change δ in bandwidth, one may take

$$\mu_2(t) = \sqrt{1 + \delta}\, \mu_1(t + \delta t), \tag{58}$$

where the factor $\sqrt{1 + \delta}$ is included to satisfy the normalization condition. A difference Ω in center frequency may be denoted by

$$\mu_2(t) = e^{i\,\Omega t}\mu_1(t), \tag{59}$$

while a difference in delay is represented by

$$\mu_2(t) = \mu_1(t + \tau). \tag{60}$$

The corresponding expressions for γ_0 are

$$\gamma_0 = B(\delta) = \sqrt{1+\delta}\int_0^\infty \mu^*(t)\mu(t+\delta t)\,dt, \qquad (61)$$

$$\gamma_0 = C(\Omega) = \int_0^\infty |\,\mu(t)\,|^2 e^{j\Omega t}\,dt,$$

$$\gamma_0 = D(\tau) = \int_0^\infty \mu^*(t)\mu(t+\tau)\,dt,$$

where the subscript 1 has been dropped. It is interesting to note that $\gamma_0(\tau)$ is just the autocorrelation function of $\mu(t)$. If one considers that the channel filters differ both in delay and center frequency, then

$$\gamma_0(\Omega, \tau) = \int_0^\infty \mu^*(t)\mu(t+\tau)e^{j\Omega t}\,dt, \qquad (62)$$

which is just Woodward's ambiguity function[10] in delay and Doppler shift for a radar pulse $\mu(t)$.

For small δ, Ω, and τ, one may obtain Taylor series expansion of $|\,\gamma_0\,|^2$ in δ, Ω, and τ [assuming certain regularity conditions on $\mu(t)$] whose first terms represent simple approximate expressions for self-noise output. In this connection, let

$$\mu(t) = a(t)e^{j\lambda(t)}, \qquad (63)$$

where $a(t)$ is the magnitude and $\lambda(t)$ is the angle of $\mu(t)$. Then it may be determined that

$$|\,B(\delta)\,|^2 = 1 - \delta^2\left\{\int_0^\infty t^2 a[a\ddot{\lambda}^2 - \ddot{a}]\,dt - \left[\int_0^\infty ta^2\dot{\lambda}\,dt\right]^2 + \frac{3}{4}\right\} + \cdots$$

$$|\,C(\Omega)\,|^2 = 1 - \Omega^2\Delta^2 + \cdots \qquad (64)$$

$$|\,D(\tau)\,|^2 = 1 - \tau^2(2\pi V)^2,$$

where Δ^2, V^2, are the mean squared duration and mean squared bandwidth, respectively, of $\mu(t)$ (recall that $|\,\mu(t)\,|^2$ and $|\,V(f)\,|^2$ have unit area),

$$\Delta = \int (t-\bar{t})^2\,|\,\mu(t)\,|^2\,dt, \qquad \bar{t} = \int t\,|\,\mu(t)\,|^2\,dt,$$

$$V = \int (f-\bar{f})\,|\,U(f)\,|^2\,df, \qquad \bar{f} = \int f\,|\,U(f)\,|^2\,df. \qquad (65)$$

The dots in (64) indicate differentiation.

Thus, the squares of the average magnitude of self noise for the three cases above are

$$(|\,\bar{N}\,|)^2 = \delta^2 X^2 \quad \text{(percentage difference of bandwidths } \delta),$$

$$(\overline{|\,N\,|})^2 = \Omega^2\Delta^2 \quad \text{(difference of center frequencies } \Omega \text{ radians/second}), \qquad (66)$$

$$(\overline{|\,N\,|})^2 = \tau^2(2\pi V)^2 \quad \text{(difference of delay } \tau).$$

[10] Woodward, *op cit.*, p. 120.

One may consider the general situation in which these three types of filter differences exist simultaneously. This problem is involved algebraically due to the appearance of six cross product terms, and no such generalizations will be presented here.

VI. Evaluation of Output Signal and Noise

The degree of noise and the amount of signal distortion present at the phase detector output will be studied in this section. It will be assumed, as in the previous section, that the noise carrier bandwidth is large compared to bandwidths of the channel filters. In this case, $\hat{\gamma}_t$ may be represented as the response of the equivalent linear filter to $e^{j\varphi(t)}$, as shown in Fig. 4. This filter has an impulse response $\mu_1^*(t)\mu_2(t)$ for the wide-band noise carrier case. Because of this representation, the problem of the determination of γ_t is reduced to a familiar PM problem: the determination of the envelope and phase of the carrier at the output of a linear narrow-band filter, when the input is a phase modulated wave. Thus, we may use the Carson-Fry[11] and Van der Pol-Stumpers[12,13] expansions. These expansions have recently been studied by Baghdady[14], who has precisely defined the conditions under which quasi-stationarity holds. The reader is referred to this paper for a discussion of the conditions leading to the validity of the quasi-stationarity representation. We will only present a heuristic approach here. Quasi-stationarity implies that the output of the filter may be computed as if the input phase derivative varied at an infinitesimal rate. Thus, in the convolution integral (40) defining γ_t, one may use the approximation

$$e^{j\varphi(t-\tau)} = e^{+j[\varphi(t)-\tau\dot{\varphi}(t)]}, \qquad (67)$$

arrived at by replacing $\varphi(t-\tau)$ by the first two terms in a Taylor series expansion about $\tau = 0$. It follows that

$$\gamma_t = e^{j\varphi(t)}\int_0^\infty \mu_1^*(\tau)\mu_2(\tau)e^{-j\tau\dot{\varphi}(t)}\,d\tau \qquad (68)$$

is the quasi-stationary representation of the complex correlation coefficient.

From the quantity $d(t)$, defined by

$$d(t) = \gamma_t e^{-j\varphi(t)},$$

it is readily seen that both the distortion of the output signal and the degree of output noise may be determined. Assuming quasi-stationarity, the angle of $d(t)$ is just the signal distortion, while $(1 - |\,d(t)\,|^2)/(|\,d(t)\,|^2)$ is the square of the average magnitude of the self-noise output.

[11] J. R. Carson and T. C. Fry, "Variable frequency electric circuit theory," *Bell Sys. Tech. J.*, vol. 16, pp. 513–540; October, 1937.

[12] B. Van der Pol, "The fundamental principles of frequency modulation," *J. IEE*, vol. 93, pt. 3, pp. 153–158; May, 1946.

[13] F. L. M. Stumpers, "Distortion of frequency modulated signals in electrical networks," *Comm. News*, vol. 9, pp. 82–92; April, 1948.

[14] E. J. Baghdady, "Theory of low distortion reproduction of FM signals in linear systems," IRE Trans. on Circuit Theory, vol. CT-5, pp. 202–214; September, 1958.

xamination of (68) shows that $d(t)$ may be interpreted the case of quasi-stationarity as the complex correlaon coefficient resulting from a separation of the channel ter center frequencies by an amount $\Omega = -\dot{\varphi}(t)$. Thus, hen quasi-stationarity is valid, one may visualize the odulation process as slowly shifting the center frequency filter 2 back and forth relative to filter 1 by an amount ual to the input frequency modulation. Since an offset center frequencies produces output noise, it is clear at the modulation process will produce a time varying mponent of output noise level in addition to whatever lf noise is present due to filter dissimilarity. When the ters are identical, it is clear from (68) and (61) that the tput signal distortion $\varphi_0(t) - \varphi(t)$ is given by

$$\varphi_0(t) - \varphi(t) = \text{angle of } C(-\dot{\varphi}). \tag{69}$$

For sufficiently small $\dot{\varphi}(t)$, the output squared average agnitude noise is

$$\overline{(|N_t|)^2} = \Delta^2(\dot{\varphi})^2. \tag{70}$$

he ensemble noise level is time varying and may be veraged to obtain a measure of the time average background noise level. The measure of this background noise level will be taken as the square root of the time average of $\overline{(|N_t|)^2}$. From (70) this is given by

$$\sqrt{\langle \overline{(|N_t|)^2} \rangle} = \Delta \sqrt{\langle (\varphi)^2 \rangle}, \tag{71}$$

where $\langle\ \rangle$ denotes the time average. Thus, in the wideband noise carrier case, the time average background self-noise level in radians (small self-noise case) is directly proportional to the RMS value of the input FM in radians/second. The proportionality constant is the RMS duration of the filter's impulse response in seconds.

In conclusion, it should be pointed out that the entire analysis of this paper has assumed the existence of a very wide bandwidth at the phase-detector output. In actuality, this bandwidth would be limited to the bandwidth of $\varphi(t)$. Thus, the noise levels predicted will only be upper bounds if the output noise bandwidth is larger than the bandwidth of $\varphi(t)$. To account for the effects of an output filter would require the evaluation of the autocorrelation function (or power spectrum) of the phase difference process $\xi(t)$. This is beyond the scope of the present paper.

Minimum-Redundancy Coding for the Discrete Noiseless Channel*

RICHARD M. KARP†, MEMBER, IRE

Summary—**This paper gives a method for constructing minimum-edundancy prefix codes for the general discrete noiseless channel ithout constraints. The costs of code letters need not be equal, and ie symbols encoded are not assumed to be equally probable. A olution had previously been given by Huffman [10] in 1952 for ie special case in which all code letters are of equal cost. The resent development is algebraic. First, structure functions are efined, in terms of which necessary and sufficient conditions for ie existence of prefix codes may be stated. From these conditions, near inequalities are derived which may be used to characterize refix codes. Gomory's integer programming algorithm is then sed to construct optimum codes subject to these inequalities; omputational experience is presented to demonstrate the practicbility of the method. Finally, some additional coding problems are iscussed and a problem of classification is treated.**

* Received by the PGIT, May 17, 1960.
† IBM Corp., Res. Center, Yorktown Heights, N. Y.

ALTHOUGH Shannon's Fundamental Theorem provides sharp upper bounds on the transmission rates achievable using letter-by-letter coding for discrete noiseless channels, the problem of constructing codes which maximize the transmission rate in specific situations has not been completely solved. Shannon [19] in 1948 and Fano [5] in 1949 gave essentially identical methods for constructing near-optimum codes when all code letters have equal cost. In 1952 Huffman [10] used an elegant combinatorial approach to obtain a strictly optimum solution to this problem. Blachman [1] in 1954 generalized the Shannon-Fano approximate technique to treat the situation in which the code letters differ in cost. Marcus [14] in 1957 improved on the Blachman technique by combining it with further combinatorial results of Huffman.

The present paper describes an algebraic approach to the construction of minimum-redundancy (minimum-cost-per-bit) prefix codes for discrete noiseless channels such that all sequences of code letters are permitted. In Section II structure functions are defined, in terms of which necessary and sufficient conditions for the existence of prefix codes may be stated. In Section III two sets of linear inequalities in integer-valued variables are used to characterize prefix codes. Next, in Section IV, Gomory's integer programming algorithm is used to construct optimum codes subject to these inequalities; computational experience is presented to demonstrate the practicability of the method. Finally, in Section V, some additional coding problems are discussed, and the results of earlier sections are applied to a problem of classification.

I. Definitions

We assume as given an alphabet of symbols $\mathfrak{A} = \{X_1, X_2, \cdots, X_n\}$ and an information source generating *source messages* consisting of sequences of these symbols in which X_i occurs with probability p_i. These messages are to be encoded for transmission over a communication channel admitting the *code letters* $s_1, s_2, \cdots, s_r$. This is to be done by associating with each element X_i of $\mathfrak{A}$ a *code word* C_i consisting of a nonempty sequence of code letters, and substituting for each source message an *encoded message* obtained by systematically substituting code words for elements of $\mathfrak{A}$.

A set $\Gamma = \{C_1, C_2, \cdots, C_n\}$ of code words is called a *code*. A code is said to be *uniquely decipherable* if each finite encoded message is an encoding of a unique source message. Sardinas and Patterson [17] in 1953 have given a test for unique decipherability.

A sequence α of code letters is a *prefix* of C_i if $C_i = \alpha \cdot \beta$, where $\cdot$ denotes juxtaposition and either α or β may be the null sequence ϕ. The prefix is called *proper* if $\beta \neq \phi$. Thus, the prefixes of the code word $s_2s_3s_1s_1s_2$ are ϕs_2, s_2s_3, $s_2s_3s_1$, $s_2s_3s_1s_1$, and $s_2s_3s_1s_1s_2$. Only the last of these is improper. A *prefix set* is the set of all elements of $\mathfrak{A}$ for which the associated code words begin with a given prefix. A *prefix code* is a code such that no code word is a prefix of any other. Any prefix code is uniquely decipherable. A prefix code is *exhaustive* if, for any two code letters s_j and s_k, $\alpha \cdot s_j$ is a prefix of a code word if and only if $\alpha \cdot s_k$ is a prefix of a code word. When an exhaustive prefix code is used, any sequence of code letters is the beginning of some encoded message.

The structure of a prefix code may be described by a tree, as shown in Fig. 1. Each terminal node is associated with an element X_i of $\mathfrak{A}$. The branches leaving each node are labeled with the names of distinct code letters, and the code word C_i is found by listing in order the labels of the branches leading from the root of the tree to the terminal node associated with X_i. The paths through any given node correspond to the symbols in a prefix set. The code is exhaustive if there are r branches leaving each node, such that each label s_k, $1 \leq k \leq r$, appears on exactly one branch. In the code of Fig. 1,

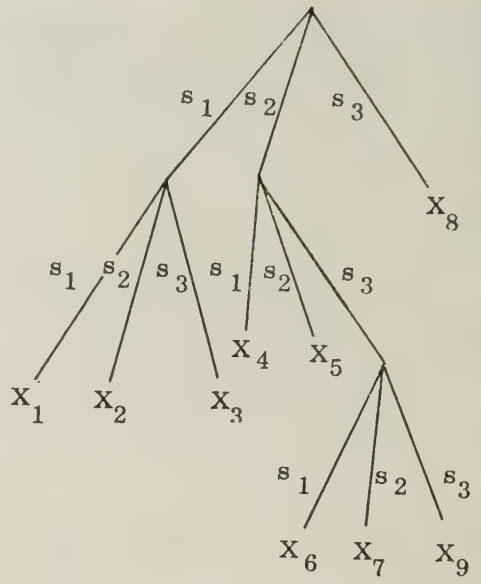


Fig. 1—Tree for an exhaustive prefix code.

$$C_1 = s_1s_1, \quad C_4 = s_2s_1, \quad C_7 = s_2s_3s_2,$$
$$C_2 = s_1s_2, \quad C_5 = s_2s_2, \quad C_8 = s_3,$$
$$C_3 = s_1s_3, \quad C_6 = s_2s_3s_1, \quad C_9 = s_2s_3s_3.$$

Let c_k be the cost of transmitting the code letter s_k. Then if N_i^k instances of s_k occur in C_i, the code word for X_i, the cost of transmitting X_i is assumed to be

$$l_i = \sum_{k=1}^{r} c_k N_i^k, \tag{1}$$

and the average cost per symbol transmitted is given by

$$\mathfrak{C} = \sum_{i=1}^{n} p_i l_i. \tag{2}$$

For uniquely decipherable codes, Shannon's Fundamental Theorem[1] yields the following lower bound:

$$\mathfrak{C} \geq \frac{-\sum_{i=1}^{n} p_i \log_2 p_i}{-\log_2 t} = \sum_{i=1}^{n} p_i \log_t p_i, \tag{3}$$

where t is the unique positive root of the equation

$$z^{c_1} + z^{c_2} + \cdots + z^{c_r} - 1 = 0. \tag{4}$$

For all communication channels, $0 \leq t \leq 1$. The quantity $-\log_2 t$ is called the *capacity* of the channel. The bound is achieved when $l_i = \log_t p_i$ $(i = 1, 2, \cdots, n)$.

The *efficiency* of a code is given by the ratio

$$E = \frac{\sum_{i=1}^{n} p_i \log_t p_i}{\mathfrak{C}}, \tag{5}$$

and the *redundancy* R is given by $R = 1 - E$.

[1] Shannon, [19], Theorem 9.

The problem of minimum-redundancy coding is to find, r a given set of probabilities p_i and costs c_k, a uniquely ecipherable code which minimizes $\mathcal{C}$. It is known [13], 6] that when all the c_k are equal, the set of optimum des must include at least one prefix code. It is not known hether this property holds in general. In this paper, tention will be restricted to the prefix codes, and a ethod will be given for minimizing $\mathcal{C}$ over this class of des. We begin by giving a structural characterization prefix codes.

II. Structure Functions for Prefix Codes

Several writers [19], [11], [13], [16] have shown that r code letters of unit cost are available, there exists a iquely decipherable code, and also a prefix code, such at C_i has cost l_i, if and only if $\sum_{i=1}^{n} r^{-l_i} \leq 1$. In the resent section, we extend this result by giving necessary nd sufficient conditions for the existence of prefix codes the general case.

As before, N_i^k will denote the number of instances s_k occurring in C_i, for a given code Γ. With each code we shall associate a polynomial form $\psi(\Gamma)$ in the vari-bles $w_1, w_2, \cdots, w_r$, which we shall call the *multivariate ructure function* (*MSF*) of Γ. This form is defined as llows:

$$\psi(C_i) = \prod_{k=1}^{r} w_k^{N_i^k};$$

$$\text{if} \quad \Gamma = (C_1, C_2, \cdots, C_n), \tag{6}$$

$$\psi(\Gamma) = \sum_{i=1}^{n} \psi(C_i) = P(w_1, w_2, \cdots, w_r).$$

ote that $\psi(\Gamma)$ does not in general have a unique inverse, e., there will exist codes Γ and Γ' such that $\psi(\Gamma) = \psi(\Gamma')$. The following theorem characterizes the exhaustive refix codes.[2]

heorem 1

The polynomial $P(w_1, w_2, \cdots, w_r)$ is the MSF of an haustive prefix code if and only if

1) $P(0, 0, \cdots, 0) = 0$,
2) all coefficients of P are nonnegative,
3) $P(w_1, w_2, \cdots, w_r) - 1 = (w_1 + w_2 + \cdots + w_r - 1) \cdot Q(w_1, w_2, \cdots, w_r)$,

here Q is a polynomial having only nonnegative co-ficients.

Before proving the theorem, we give some examples r the case $r = 2$:

a) $P = 2w_1 + 2w_2 + 3$;

condition 1) is violated.

b) $P = w_1^3 + w_1^2 w_2 - w_1^2 + w_1 + w_2$;

condition 2) is violated.

c) $P = w_1^3 + w_2^3 + 3w_1w_2$; $P - 1 = (w_1 + w_2 - 1) \cdot (w_1^2 + w_2^2 - w_1w_2 + w_1 + w_2 + 1)$;

condition 3) is violated.

d) $P = w_1^3 + w_2^3 + w_1w_2^2 + w_1w_2 + w_1$;

$P - 1$ is not divisible by $(w_1 + w_2 - 1)$;

condition 3) is violated.

e) $P = w_2^3 + w_1w_2^2 + w_1w_2 + w_1$;

$(P - 1) = (w_1 + w_2 - 1)(w_2^2 + w_2 + 1)$;

$P = \psi(\Gamma)$, where $\Gamma = \{s_2s_2s_1, s_2s_2s_2, s_2s_1, s_1\}$.

Proof:

Necessity: If $P(w_1, w_2, \cdots, w_r)$ is the MSF of an exhaustive prefix code, then it satisfies 1) and 2) by definition. To show that 3) is satisfied, we use induction on the maximum code-word length, where the length of C_i is $\sum_{k=1}^{r} N_i^k$. The only exhaustive prefix code with maximum length 1 has the MSF $w_1 + w_2 + \cdots + w_r$, which satisfies 3) with $Q = 1$. Suppose 3) holds for all exhaustive prefix codes of maximum length less than L. Let $P(w_1, w_2, \cdots, w_r)$ be the MSF of an exhaustive prefix code of maximum length L, and let $S(w_1, w_2, \cdots, w_r)$ be the sum of the terms of degree L. Then, by the definition of exhaustiveness, for all code letters s_i and s_k, $\alpha \cdot s_i$ is a code word of length L if and only if $\alpha \cdot s_k$ is a code word of length L. Therefore, S may be written as

$$S(w_1, w_2, \cdots, w_r) = (w_1 + w_2 + \cdots + w_r)(\psi(\alpha_1) + \psi(\alpha_2) + \cdots + \psi(\alpha_p))$$

where $\alpha_1, \alpha_2, \cdots, \alpha_p$ are the proper prefixes of length $L - 1$ in the code Γ. Let $\psi(\alpha_1) + \psi(\alpha_2) + \cdots + \psi(\alpha_p)$ be denoted by $T(w_1, w_2, \cdots, w_r)$, where all the terms of T have positive coefficients. Then the function P may be expressed as

$$P(w_1, w_2, \cdots, w_r) = (w_1 + w_2 + \cdots + w_r) \cdot T(w_1, w_2, \cdots, w_r) + V(w_1, w_2, \cdots, w_r)$$

where V is of degree less than L and has only nonnegative coefficients. Then if the rp code words of length L are replaced by the p prefixes $\alpha_1, \alpha_2, \cdots, \alpha_p$, a new exhaustive prefix code is obtained, having the MSF

$$\begin{aligned}\bar{P}(w_1, w_2, \cdots, w_r) &= P(w_1, w_2, \cdots, w_r) \\ &\quad - (w_1 + w_2 + \cdots + w_r)T(w_1, w_2, \cdots, w_r) \\ &\quad + T(w_1, w_2, \cdots, w_r) \\ &= T(w_1, w_2, \cdots, w_r) + V(w_1, w_2, \cdots, w_r),\end{aligned}$$

and by the induction hypothesis,

$$\bar{P}(w_1, w_2, \cdots, w_r) - 1 = (w_1 + w_2 + \cdots + w_r - 1)\bar{Q}(w_1, w_2, \cdots, w_r)$$

[2] In the course of developing this theorem, the writer benefited om discussions with Profs. D. E. Muller and M. P. Schützen-rger.

where $\bar{Q}$ has only nonnegative coefficients. Therefore,

$$P(w_1, w_2, \cdots, w_r) - 1 = (w_1 + w_2 + \cdots + w_r - 1) \cdot (T(w_1, w_2, \cdots, w_r) + \bar{Q}(w_1, w_2, \cdots, w_r)),$$

and (3) is satisfied.

Sufficiency: We shall show that, if P satisfies 1), 2), and 3), then it is the MSF of an exhaustive prefix code. Let P be of degree 1. Substituting $w_1 = w_2 = \cdots = w_r = 0$ in 3), and observing that 1) holds, we find that

$$-1 = (-1)Q(0, 0, \cdots, 0).$$

Therefore $Q(0, 0, \cdots, 0) = 1$; also, Q is of degree zero, since P is of degree 1. Therefore $Q = 1$ and $P = w_1 + w_2 + \cdots + w_r$, which is the MSF of an exhaustive prefix code. We now proceed by induction. Suppose the result holds for all polynomials of degree less than L, and let P be of degree L. Since 3) holds,

$$P(w_1, w_2, \cdots, w_r) - 1 = (w_1 + w_2 + \cdots + w_r - 1) \cdot (T(w_1, w_2, \cdots, w_r) + \bar{Q}(w_1, w_2, \cdots, w_r))$$

where T is homogeneous of degree $L - 1$, and all terms of $\bar{Q}$ are of degree less than $L - 1$; $\bar{Q}$ and T contain only nonnegative coefficients. Then

$$\begin{aligned}\bar{P}(w_1, w_2, \cdots, w_r) &= P(w_1, w_2, \cdots, w_r) \\ &\quad - (w_1 + w_2 + \cdots + w_r)T(w_1, w_2, \cdots, w_r) \\ &\quad + T(w_1, w_2, \cdots, w_r) \qquad (7)\end{aligned}$$

satisfies 1), since $P(0, 0, \cdots, 0) = T(0, 0, \cdots, 0) = 0$. Also, since

$$\bar{P}(w_1, w_2, \cdots, w_r) = (w_1 + w_2 + \cdots + w_r - 1) \cdot \bar{Q}(w_1, w_2, \cdots, w_r), \qquad (8)$$

3) is satisfied. To verify 2), note that, since P and T have only nonnegative coefficients, the only possible negative terms in $\bar{P}$ arise from the expression

$$-(w_1 + w_2 + \cdots + w_r)T(w_1, w_2, \cdots, w_r),$$

and are of degree L. But (8), in which $\bar{Q}$ is of degree less than $L - 1$, shows that no negative terms of degree L can occur in $\bar{P}$. Therefore, by the induction hypothesis, $\bar{P}$ is the MSF of an exhaustive prefix code. Moreover, by inspection of (7), each term of T appears as a term in $\bar{P}$ (if $\bar{P}$ is expressed with only unit coefficients), since no term of T is canceled by a term in P, which has only nonnegative coefficients, or by a term in $-(w_1 + w_2 + \cdots + w_r)T$, which is homogeneous of degree L. Then each term in T is $\psi(\alpha)$ for some code word α of length $L - 1$ in the code associated with $\bar{P}$. If each such code word is replaced by the r code words $\alpha \cdot s_1, \alpha \cdot s_2, \cdots, \alpha \cdot s_r$, an exhaustive prefix code having the MSF P is exhibited.

The quotient polynomial Q may be characterized as follows:

Corollary 1.1: If $P = \psi(\Gamma)$, where Γ is an exhaustive prefix code whose code words have the distinct proper prefixes $\beta_1, \beta_2, \cdots, \beta_q$, and

$$(P - 1) = (w_1 + w_2 + \cdots + w_r - 1)Q,$$

then $Q = \psi(\beta_1) + \psi(\beta_2) + \cdots + \psi(\beta_q)$.

Proof: Again we use induction on the maximum code word length L. If $L = 1$, $P = w_1 + w_2 + \cdots + w_r$, $Q = 1 = \psi(\phi)$, and the result holds. Suppose the theorem is true for codes having maximum word length L. Let P be $\psi(\Gamma)$, where Γ has maximum word length L, and define Q, T, $\bar{P}$, and $\bar{Q}$ as in the proof of Theorem I. Then $\bar{P} = \psi(\bar{\Gamma})$, where $\bar{\Gamma}$ has the same proper prefixes as Γ, except for those of length $L - 1$, which no longer occur. By the induction hypothesis,

$$\bar{Q} = \psi(\gamma_1) + \psi(\gamma_2) + \cdots + \psi(\gamma_q),$$

where $\gamma_1, \gamma_2, \cdots, \gamma_q$ are the proper prefixes of $\bar{\Gamma}$. Also, $T = \psi(\alpha_1) + \psi(\alpha_2) + \cdots + \psi(\alpha_p)$, where $\alpha_1, \alpha_2, \cdots, \alpha_p$ are the proper prefixes of Γ having length $L - 1$. Therefore, $Q = T + \bar{Q}$ has the required property.

Example 1: In the code corresponding to the tree of Fig. 1, we see by inspection that

$$\begin{aligned}P &= \psi(s_1s_1) + \psi(s_1s_2) + \psi(s_1s_3) + \psi(s_2s_1) + \psi(s_2s_2) \\ &\quad + \psi(s_2s_3s_1) + \psi(s_2s_3s_2) + \psi(s_3) + \psi(s_2s_3s_3) = w_2w_3^2 \\ &\quad + w_2^2w_3 + w_1w_2w_3 + w_1^2 + 2w_1w_2 + w_1w_3 + w_2^2 + w_3.\end{aligned}$$

Inspecting the nonterminal nodes of the tree, which correspond to proper prefixes, we find that

$$\begin{aligned}Q &= \psi(\phi) + \psi(s_1) + \psi(s_2) + \psi(s_2s_3) \\ &= w_2w_3 + w_1 + w_2 + 1,\end{aligned}$$

and

$$P - 1 = (w_1 + w_2 + w_3 - 1)Q.$$

The following method may be used to construct an exhaustive prefix code Γ having a given MSF P:

1) Define a polynomial $R(w_1, w_2, \cdots, w_r)$, initially equal to $Q(w_1, w_2, \cdots, w_r)$, and a set S of sequences of code letters, initially equal to $\{\phi\}$.
2) Choose an element $\beta \in S$, not previously selected. If $\psi(\beta)$ is a term in R, replace β by the r sequences $\beta \cdot s_1, \beta \cdot s_2, \cdots, \beta \cdot s_r$, and replace R by $R - \psi(\beta)$. If $\psi(\beta)$ is not a term in R, select another element.
3) Continue until $R \equiv 0$. At this point, S will be the desired code.

Next, we give a characterization of prefix codes in terms of the costs of code words. For this purpose, we assume that the costs of the code letters are integers; of course, multiplication by a suitable constant suffices to transform any set of rational costs into integers, without changing the problem of constructing optimum codes. As before, let c_k denote the cost of s_k, and let N_i^k denote the number of occurrences of s_k in C_i, for a code Γ. We assume without loss of generality that $c_1 \leq c_2 \leq \cdots \leq c_r$.

e cost of C_i is given by

$$l_i = \sum_{k=1}^{r} c_k N_i^k.$$

define $F(z) = \sum_{i=1}^{n} z^{l_i}$ as the *univariate structure ction (USF)* of Γ.

orem 2

The polynomial $F(z) = \sum_{j=1}^{m} a_j z^j$ is the USF of an austive prefix code if and only if

) $a_j \geq 0$ for all j,
) $F(z) - 1 = z^{c_1} + z^{c_2} + \cdots + z^{c_r} - 1)G(z)$,

where $G(z) = \sum_{j=0}^{m-c_r} b_j z^j$, and $b_j \geq 0$ for all j.

Proof: The result that the USF of any exhaustive pre- code satisfies 1) and 2) is obtained by making the stitution $w_k = z^{c_k}$ in condition 3) of Theorem I. The verse is proved by induction on the degree of $F(z)$; steps in the proof correspond exactly to those in the of of sufficiency in Theorem I.

Corollary 2.1: The polynomial $\sum_{j=1}^{m} a_j z^j$ is the USF a prefix code if and only is there exist integers d_j such at

) $a_j \leq d_j$, $1 \leq j \leq m$,
) $\sum_{j=1}^{m} d_j z^j$ is the USF of an exhaustive prefix code.

Corollary 2.2: If $F(z)$ is the USF of an exhaustive prefix le Γ, and $F(z) - 1 = (z^{c_1} + z^{c_2} + \cdots + z^{c_r} - 1)G(z)$, ere $G(z) = \sum_{j=0}^{m} b_j z^j$, then b_j is the number of proper efixes of Γ having cost j.

Example 2:

a) $c_1 = 1$, $c_2 = 2$.

$F(z) = z^6 + 4z^3$; $F(z) - 1 = (z^2 + z - 1)(z^4 - z^3 + 2z^2 + z + 1)$; $G(z)$ has a negative coefficient, violating 2).

b) $c_1 = 1$, $c_2 = c_3 = 2$.
$F(z) = 2z^5 + 3z^4 + 2z^3 + 2z^2$; $F(z) - 1 = (2z^2 + z - 1)(z^3 + z^2 + z + 1)$; $F(z)$ is the USF of an exhaustive prefix code.

c) $c_1 = 1$, $c_2 = c_3 = 2$.
$F(z) = z^5 + 3z^4 + 2z^3 + 2z^2$; comparison with b) shows that $F(z)$ is the USF of a prefix code.

If β is the sequence $s_{i_1}, s_{i_2}, \cdots, s_{i_p}$ we define $\Phi(\beta) = {}_{1}+c_{i_2}+\ldots+c_{i_p}$. Using this notation, we may give the follow- rule for constructing a code having a given USF $F(z)$.

1) Define a set S, initially equal to $\{\phi\}$, and a polynomial $H(z)$, initally equal to $F(z)$.
2) Choose an element $\beta \in S$, not previously selected. If $\Phi(\beta)$ is a term in $H(z)$, replace $H(z)$ by $H(z) - \Phi(\beta)$; if not, replace β by the r sequences $\beta \cdot s_1$, $\beta \cdot s_2$, $\cdots$, $\beta \cdot s_r$.
3) Continue until $H(z) \equiv 0$; S will then represent a code having the USF $\hat{F}(z)$.
4) For each term z^l in $\hat{F}(z) - F(z)$, delete an element β of S such that $\Phi(\beta) = z^l$.

III. Linear Constraints on Prefix Codes

In this section we derive certain linear inequalities in the coefficients a_j and b_j. Interpretations of these inequalities in terms of the structures of codes will be given. In the following section, the inequalities will be used to express the construction of optimum codes as a problem amenable to solution by Gomory's integer programming algorithm [7]–[9].

As before, let $F(z) = \sum_{j=1}^{m} a_j z^j$, and $G(z) = \sum_{j=0}^{m-c_r} b_j z^j$. Then for exhaustive prefix codes,

$$F(z) - 1 = (z^{c_1} + \cdots + z^{c_r} - 1)G(z),$$

and, equating the coefficients of z^j on both sides of the equation, we find that $a_0 = 0$, $b_0 = 1$, and,

for $j > 0$,

$$a_j = \sum_{k=1}^{r} b_{j-c_k} - b_j, \tag{9}$$

with the convention that $b_j = 0$, $j < 0$. For prefix codes in general, Corollary 2.1 implies that (9) may be reduced to an inequality,

$$a_j \leq \sum_{k=1}^{r} b_{j-c_k} - b_j. \tag{10}$$

Thus, the results of the previous section may be restated as follows.

Theorem 3

There exists a prefix code having a_j code words of cost j if and only if there exist nonnegative integers b_j such that,

$$\text{for } j > 0, \quad a_j \leq \sum_{k=1}^{r} b_{j-c_k} - b_j, \tag{a}$$

$$\text{for } j > 0, \quad a_j \geq 0, \tag{b}$$

$$a_0 = 0;\ b_0 = 1;\ a_j = b_j = 0,\ j < 0. \tag{c}$$

The code is exhaustive if and only if equalities hold in (a).

These conditions may be given a simple interpretation by noting that, for an exhaustive prefix code, the number of prefixes of length j terminating in s_k is b_{j-c_k}. Summing over all k, the number of prefixes of length j is $\sum_{k=1}^{r} b_{j-c_k}$. Each of these is either a proper prefix, of which there are b_j, or a code word, of which there are a_j. Thus, we obtain the identity

$$a_j + b_j = \sum_{k=1}^{r} b_{j-c_k}. \tag{11}$$

For a nonexhaustive code, some sequences are unused, and (11) may be weakened to an inequality. Thus, the conditions of Theorem 3 are shown to be necessary for a prefix code; their sufficiency is established by Theorem 2 and Corollary 2.1.

Another interesting form of these inequalities is obtained by expressing the b_j in terms of the a_j. To assist in this we use matrix notation. We begin by considering only exhaustive prefix codes. Let the column vectors A and B be defined as follows:

$$A = (-1, a_1, a_2, \cdots, a_m)^T,$$
$$B = (1, b_1, b_2, \cdots, b_{m-c_r})^T.$$

Let $P(x)$ denote the number of code letters of cost x, and let the $(m + 1) \times (m + 1 - c_r)$ matrix M be defined as

$$(M)_{ij} = \begin{cases} -1, & i - j = 0 \\ P(i - j), & \text{otherwise.} \end{cases}$$

Then, according to Theorem 3,

$$MB = A. \tag{12}$$

Example 3:

$$c_1 = 1, \quad c_2 = c_3 = 2,$$
$$F(z) = 2z^5 + 3z^4 + 2z^3 + 2z^2, \quad G(z) = z^3 + z^2 + z + 1,$$
$$A = (-1, 0, 2, 2, 3, 2)^T, \qquad B = (1, 1, 1, 1)^T,$$

$$\begin{matrix} & M & & & & B & = & A \end{matrix}$$
$$\begin{bmatrix} -1 & 0 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 2 & 1 & -1 & 0 \\ 0 & 2 & 1 & -1 \\ 0 & 0 & 2 & 1 \\ 0 & 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \end{bmatrix} = \begin{bmatrix} -1 \\ 0 \\ 2 \\ 2 \\ 3 \\ 2 \end{bmatrix}.$$

Let $\bar{M}$ denote the square matrix consisting of the first $m + 1 - c_r$ rows of M, and let $\bar{A}$ denote the vector consisting of the first $m + 1 - c_r$ elements of A. Then

$$\bar{M}B = \bar{A} \tag{13}$$

and, if $\bar{M}$ is nonsingular,

$$\bar{M}^{-1}\bar{A} = B \geq 0. \tag{14}$$

We shall exhibit $\bar{M}^{-1}$. Let the following recursive sequence be defined:

$$v_h = \begin{cases} 0, & h < 0 \\ 1, & h = 0 \\ \sum_{k=0}^{r} v_{h-c_k}, & h > 0. \end{cases}$$

Thus, for $c_1 = 1, c_2 = 2$, the following sequence is defined, beginning at $h = 0$:

$$1, 1, 2, 3, 5, 8, 13, \cdots.$$

In general, v_h is the number of possible sequences of code letters having cost h.

Lemma 4.1:

$$(\bar{M}^{-1})_{ij} = -v_{i-j}.$$

Proof: Let H be an $(m + 1 \times c_r) - (m + 1 - c_r)$ matrix such that $(H)_{ij} = -v_{i-j}$. We must show that

$$(\bar{M}H)_{ij} = \sum_{q=1}^{m+1-c_r} (\bar{M})_{iq}H_{qj} = \delta_i^j.$$

Using the definitions of $\bar{M}$ and H,

$$(\bar{M}H)_{ij} = -H_{ij} + \sum_{k=1}^{r} H_{i-c_k,j} = v_{i-j} - \sum_{k=1}^{r} v_{i-j-c_k}. \tag{16}$$

Case 1: $i - j < 0$. All the terms in (16) are zero, and $(\bar{M}H)_{ij} = 0$.

Case 2: $i - j = 0$. All the terms except v_{i-j} are zero; $(\bar{M}H)_{ii} = v_0 = 1$.

Case 3: $i - j > 0$. By (15), the value of (16) is zero.

Thus, (14) is equivalent to $(m + 1 - c_r)$ linear inequalities, the $(j + 1)$st of which is

$$v_j - \sum_{p=1}^{j} v_{j-p}a_p = b_j \geq 0. \tag{17}$$

These conditions are equivalent to the first $(m + 1 - c_r)$ equations implied by (12). The $(s + 1)$st equation, $s > m - c_r$, is

$$a_s = \sum_{k=1}^{r} b_{s-c_k}. \tag{18}$$

Substituting (17) in (18), we find that

$$a_s = \sum_{k=1}^{r} v_{s-c_k} - \sum_{k=1}^{r} \sum_{p=1}^{s-c_k} v_{s-c_k-p}a_p$$
$$= v_s - \sum_{k=1}^{r} \sum_{p=1}^{s-c_k} v_{s-c_k-p}a_p.$$

Reversing the order of summation, we find that

$$a_s = v_s - \sum_{p=1}^{s-1} a_p \sum_{k=1}^{r} v_{s-c_k-p},$$

and by (15),

$$a_s = v_s - \sum_{p=1}^{s-1} a_p v_{s-p},$$

or

$$v_s = \sum_{p=1}^{s} a_p v_{s-p}. \tag{19}$$

The results of the foregoing calculations are summarized in Theorem 4.

Theorem 4

There exists an exhaustive prefix code with a_p words of cost p, and maximum code-word cost m, if and only if,

$$\text{for} \quad 1 \leq j \leq m - c_r, \qquad \sum_{p=1}^{j} v_{j-p}a_p \leq v_j, \tag{20}$$

$$\text{for} \quad m - c_r < j \leq m, \qquad \sum_{p=1}^{j} v_{j-p}a_p = v_j. \tag{21}$$

orollary 4.1: There exists a prefix code having a_p words ost p if and only if, for $1 \leq j \leq m$,

$$\sum_{\nu=1}^{j} v_{j-\nu} a_\nu \leq v_j. \tag{22}$$

roof: Given a prefix code Γ with a USF violating (22), ollary 2.1 asserts that the coefficients a_p may be in- sed to yield the USF of an exhaustive prefix code Γ'; the new coefficients would violate either (20) or (21), radicting Theorem 4. Therefore, (22) is a necessary dition. To prove sufficiency, define

$$\delta_j = v_j - \sum_{p=1}^{j} v_{j-p} a_p. \tag{23}$$

ng (15), we find that (23) becomes

$$\delta_j = \sum_{k=1}^{r} \delta_{j-c_k} - a_j, \tag{24}$$

for $j > m$,

$$\delta_j = \sum_{k=1}^{r} \delta_{j-c_k}. \tag{25}$$

en given a USF satisfying (22), we may increase its fficients a_j to satisfy (20) and (21) by a process of c_r os, the kth of which is as follows: increase a_{m+k-1} by set δ_m equal to zero, and recompute the δ_j, $j > m +$ 1, according to (25). When the process is completed, conditions of Theorem 4 will be satisfied (with m efined as the maximum code-word length of the newly ived code). Thus, by Corollary 2.1, the original coeffi- nts a_j correspond to a prefix code.

Ve note that, if (19) holds for $m - c_r < j \leq m$, then olds also for $q > m$, since

$$\delta_q = \sum_{k=1}^{r} \delta_{q-c_r} = 0.$$

imilar comment holds for (22).

Example 4:

Let $c_1 = 1$, $c_2 = 2$, $c_3 = 3$.
Then $v_0 = 1$, $v_1 = 1$, $v_2 = 2$, $v_3 = 4$, $v_4 = 7$, $v_5 = 13$, $_6 = 24$.
Suppose $m = 5$, $a_1 = 0$, $a_2 = 1$, $a_3 = a_4 = a_5 = 2$.

en

$$\delta_1 = v_1 - a_1 = 1,$$

$$\delta_2 = v_2 - v_1 a_1 - a_2 = 1,$$

$$\delta_3 = v_3 - v_2 a_1 - v_1 a_2 - a_3 = 1,$$

$$\delta_4 = v_4 - v_3 a_1 - v_2 a_2 - v_1 a_3 - a_4 = 1,$$

$$\delta_5 = v_5 - v_4 a_1 - v_3 a_2 - v_2 a_3 - v_1 a_4 - a_5 = 1.$$

e conditions of Corollary 4.1 are satisfied. Coefficients isfying (20) and (21) are obtained as follows: increase by 1, a_6 by 2, and a_7 by 1, giving $a_1 = 0$, $a_2 = 1$, $a_3 =$ $= 2$, $a_5 = 3$, $a_6 = 2$, $a_7 = 1$.

The condition (22) lends itself to a simple combinatorial interpretation. Since there are exactly v_{j-p} sequences of code letters having cost $j - p$, each code word of cost p is the prefix of v_{j-p} sequences of cost j. By the definition of a prefix code, no sequence can have two distinct code words as prefixes. Therefore, the number of distinct sequences of cost j having code words as prefixes is $\sum_{p=1}^{j} v_{j-p} a_p$, and this number may not exceed v_j, the total number of sequences of cost j. This argument establishes only the necessity of (23); Corollary 4.1 establishes its sufficiency as well.

Recalling the definition of t as the unique positive root of

$$z^{c_1} + z^{c_2} + \cdots + z^{c_r} = 1, \tag{26}$$

we assert that, for a uniquely decipherable code,

$$\sum_{i=1}^{n} t^{l_i} \leq 1. \tag{27}$$

For, if the contrary is assumed, let

$$\sum_{i=1}^{n} t^{l_i} = t^{-\delta}, \qquad \delta < 0.$$

Then, if $p_i = t^{l_i + \delta}$,

$$\sum_{i=1}^{n} p_i \log_t p_i = \delta + \sum_{i=1}^{n} p_i l_i,$$

contradicting Shannon's Fundamental Theorem.

It is of interest to compare (27) with the conditions of Corollary 4.1. It is well known (Lagrange [12]; see also Dickson [4]) that, for a sequence having the rule of formation given by (15), the general term v_h may be expressed in terms of the roots $t_1, t_2, \cdots, t_{c_r}$ of (26). If the roots are distinct,

$$v_h = \sum_{i=1}^{c_r} \lambda_i t_i^{-h}.$$

In the case of a repeated root t_i, having multiplicity μ, λ_i is replaced by

$$(\lambda_{i0} + \lambda_{i1} h + \cdots + \lambda_{i,\mu-1} h^{\mu-1}).$$

It can be shown, using a theorem of Cauchy quoted by Marden,[3] that t, the root of (26) having least absolute value, is unique, and is the only positive root, provided that the c_k have no common factor. Then

$$\frac{v_h}{v_{h+1}} = t\left[1 + \mathcal{O}\left(\frac{t}{u}\right)^h\right],$$

where u is the second smallest root of (26), and

$$\lim_{h \to \infty} \frac{v_h}{v_{h+1}} = t.$$

Dividing through by v_j in (22), one obtains

$$\sum_{p=1}^{j} \frac{v_{j-p}}{v_j} a_p \leq 1,$$

[3] Marden, [15], p. 95.

and, except for the "parasitic" roots of (26) (*i.e.*, those other than t), this would reduce to

$$\sum_{p=1}^{i} a_p t^p \leq 1,$$

which is implied by (27). In the case $c_r = 1$, there are no parasitic roots, and we can prove Corollary 4.2.

Corollary 4.2 (Kraft-Szilard inequality): If $c_1 = c_2 = \cdots = c_r = 1$, there exists a prefix code having a_j words of cost j, and maximum code-word length m, if and only if

$$\sum_{p=1}^{m} a_p r^{-p} \leq 1. \tag{28}$$

Proof: In this case $v_h = rv_{h-1} = r^h$, and (22) reduces to,

$$\text{for} \quad 1 \leq j \leq m, \qquad \sum_{p=1}^{j} r^{j-p} a_p \leq r^j$$

or, equivalently,

$$\text{for} \quad 1 \leq j \leq m, \qquad \sum_{p=1}^{j} r^{-p} a_p \leq 1,$$

which is equivalent to (28).

IV. An Integer Programming Method for Constructing Optimum Codes

In the previous two sections, various necessary and sufficient conditions for the existence of prefix codes have been obtained. In this section, these results are employed in the construction of optimum prefix codes. The problem may be stated as follows:

Minimize

$$\sum_{i=1}^{n} p_i l_i$$

subject to,

$$b_0 = 1; b_j = 0, j < 0$$

$$\text{for} \quad 1 \leq j \leq m, \qquad a_j + b_j \leq \sum_{k=1}^{r} b_{j-c_k} \tag{29a}$$

or, equivalently,

$$\text{for} \quad 1 \leq j \leq m, \qquad \sum_{p=1}^{j} a_p v_{j-p} \leq v_j. \tag{29b}$$

In order to make this a well-stated minimization problem, it is necessary to express the relations between the variables l_i, which give the costs of specific code words, and the a_j, which tell how many code words of a given cost occur. In the special case when $p_1 = p_2 = \cdots = p_n = 1/n$,

$$\sum_{i=1}^{n} p_i l_i = \frac{1}{n} \sum_{j=1}^{m} j a_j,$$

and (29), together with the constraint

$$\sum_{j=1}^{m} a_j = n,$$

is sufficient. More generally, we may define

$$y_{ij} = \begin{cases} 1 & \text{if} \quad l_i = j \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$l_i = \sum_{j=1}^{m} j y_{ij}, \qquad a_j = \sum_{i=1}^{n} y_{ij},$$

and the problem becomes

$$\min \sum_{i=1}^{n} \sum_{j=1}^{m} p_i j y_{ij} \tag{30}$$

subject to

$$b_0 = 1; b_j = 0, j < 0$$

$$\text{for} \quad 1 \leq j \leq m, \qquad \sum_{i=1}^{n} y_{ij} + b_j \leq \sum_{k=1}^{r} b_{j-c_k} \tag{31a}$$

or equivalently

$$\text{for} \quad 1 \leq j \leq m, \qquad \sum_{p=1}^{j} \sum_{i=1}^{n} y_{ip} v_{j-p} \leq v_j \tag{31b}$$

and

$$\text{for} \quad 1 \leq i \leq n, \qquad \sum_{j=1}^{m} y_{ij} = 1,$$

together with the requirement that all variables y_{ij} and b_j be non-negative integers.

The problem has now been cast as the minimization of a linear function subject to linear constraints, together with the requirement that the variables assume integer values. Gomory [7]–[9] has developed a computational method called integer programming for solving problems of this type.[4] In principle, this method can be applied here, once an upper bound m on the costs of code words has been set. Since mn integer variables y_{ij} are involved, however, such an approach would permit the solution of only very small problems. To extend the practical scope of the method, it is necessary to reduce the number of variables actually needed in the computation. This problem will be considered next.

Let the X_i be so ordered that, if $i < k$, $p_i \geq p_k$. Then there exists an optimum code such that $l_i \leq l_k$ whenever $i < k$; such codes will be called *monotone*. All optimum codes are monotone unless some of the p_i are equal. Let $\mathcal{C}(i, j)$ be a lower bound on the cost of a monotone code such that $y_{ij} = 1$. Suppose y_{ij} is set to zero in (30) and (31) whenever $\mathcal{C}(i, j) > \mathcal{C}_0$, and the resulting problem in a reduced number of variables is solved. Then, if the solution obtained has cost less than or equal to $\mathcal{C}_0$, it is the optimum solution to the original problem. If the $\mathcal{C}(i, j)$ are sharp lower bounds, and $\mathcal{C}_0$ is properly chosen, the number of variables in the integer programming calculation may be reduced greatly.

[4] The author is indebted to Dr. R. E. Gomory for helpful discussions of the theory and practice of integer programming, and for making his IBM 704 program available.

We recall that the condition

$$\sum_{i=1}^{n} t^{l_i} = \sum_{p=1}^{m} a_p t^p \leq 1 \tag{32}$$

necessary for the existence of a prefix code, and that constraint

$$\sum_{p=1}^{j} a_p t^p \leq 1,$$

plied by (32), closely approximates the jth linear constraint in (29b), with an exponentially vanishing error. Indeed, we have found empirically that the replacement of (29) by (32) changes the minimum value of $\sum_{i=1}^{n} p_i l_i$ little, if at all. Thus, a sharp bound $\mathcal{C}(i_0, j_0)$ can be found by solving the following problem, which will be shown to require very little computation:

$$\min \sum_{i=1}^{n} p_i l_i$$

subject to

$$\sum_{i=1}^{n} t^{l_i} \leq 1, \tag{a}$$

$$\text{for} \quad i \leq i_0, \qquad l_i \leq j_0, \tag{b}$$

$$\text{for} \quad i \geq i_0, \qquad l_i \geq j_0, \tag{c}$$

where (b) and (c) are required for monotonicity. Here, t is simply a constant which may be obtained from (26) by standard numerical procedures.

This problem can be converted to an integer programming problem by defining

$$x_{ij} = \begin{cases} 1, & j \leq l_i \\ 0, & j > l_i. \end{cases}$$

Then

$$t^{l_i} = 1 - (1 - t) \sum_{j=1}^{m} t^{j-1} x_{ij}, \qquad l_i = \sum_{j=1}^{m} x_{ij}.$$

Making these substitutions, one obtains

$$\min \sum_{i=1}^{n} \sum_{j=1}^{m} p_i x_{ij},$$

subject to

$$\sum_{i=1}^{n} \sum_{j=1}^{m} t^{j-1} x_{ij} \geq \frac{n-1}{1-t} \tag{33}$$

$$\text{for} \quad i \leq i_0, \qquad j > j_0, \qquad x_{ij} = 0 \tag{34}$$

$$\text{for} \quad i \geq i_0, \qquad j \leq j_0, \qquad x_{ij} = 1. \tag{35}$$

In order that the x_{ij} may meaningfully define the l_i, it is necessary that, whenever $j_1 < j_2$, $x_{ij_1} \geq x_{ij_2}$. If, however, a solution to the above problem had $x_{ij_1} = 0$, $x_{ij_2} = 1$, $j_2 > j_1$, a solution having lower cost could be obtained by interchanging the values of x_{ij_1} and x_{ij_2}; therefore, the optimum solution to the above problem will automatically satisfy this requirement.

Apart from upper bounds on the values of the nonnegative variables x_{ij}, and conditions which give preassigned values to the x_{ij}, (33) is the only constraint to be satisfied. Integer programming problems of this type are called knapsack problems, and are relatively easy to solve [3]. If the constraints are weakened to admit nonintegral values for the x_{ij}, the solution becomes particularly simple. The variables x_{ij} which do not have preassigned values are set equal to 1 in increasing order of the quantity $p_i t^{-j}$ until, after a variable x_{uv} has been set equal to one, (33) is satisfied. The value of x_{uv} is then decreased so that (33) is satisfied as an equality.

Example 5: Let $n = 10$, $r = 2$, $c_1 = 1$, $c_2 = 3$.

$$t^3 + t - 1 = 0; \qquad t = 0.682.$$

Let the p_i be 0.23, 0.20, 0.16, 0.11, 0.08, 0.07, 0.05, 0.04, 0.03, 0.03. We shall determine $\mathcal{C}(2, 5)$. In the matrix of Fig. 2, the i, j element corresponds to the variable x_{ij}.

i \ j	1	2	3	4	5	6	7	8	9	10	11
1	3	7	13	19							
2											
3											
4											
5						14					
6						12	18				
7						8	13	20			
8						4	9	15			
9						2	6	11	17		
10						1	5	10	16		

Fig. 2—Matrix used in computing $\mathcal{C}$ (2,5).

The dotted area corresponds to those variables having the preassigned value zero, according to (34). Variables in the cross-hatched area have the preassigned value 1, according to (35). The remaining elements are numbered in the order of their inclusion in the solution. The constraint (33) is first satisfied when x_{78} is set equal to 1; it is satisfied as an equality when $x_{78} = 0.33$. The following solution is thus obtained:

$$l_1 = 4, \quad l_2 = l_3 = l_4 = 5, \quad l_5 = 6,$$
$$l_6 = 7, \quad l_7 = 7.33, \quad l_8 = 8, \quad l_9 = l_{10} = 9,$$

$$\sum_{i=1}^{n} p_i l_i = 5.46$$

Therefore $\mathcal{C}(2, 5) = 5.46$.

Once the cells of the matrix have been arranged in increasing order of $p_i \cdot t^{-i}$, little calculation is needed to determine $\mathcal{C}(i, j)$ for any values of i and j.

When p_i is small, j can be made very large without causing a large increase in the value of $\mathcal{C}(i, j)$. However, if Δ_k is an upper bound on $l_{i+k} - l_i$, for an optimum code, we may define the following improved bound:

$$\mathcal{C}^*(i, j) = \max_{g>0} \min_{j-\Delta_i-g \le h \le j} \mathcal{C}(g, h).$$

This bound will be of value principally when i is close to n. We state the following result without proof:

If $k = (r-1)Q(k) + R(k)$, $0 \le R(k) < r-1$, then

$$\Delta_k = \max_{j=0,1,\cdots,r-2} (Q(k+j)c_r + c_{r-1-j} - c_{r-1-R(k+j)}).$$

It remains to discuss the choice of $\mathcal{C}_0$. Clearly, $\mathcal{C}_0$ should be large enough so that there exists a code of cost less than or equal to $\mathcal{C}_0$, but otherwise as small as possible. In the examples treated thus far, a near-optimum code was constructed by cut-and-try procedures, and $\mathcal{C}_0$ was taken as the cost of that code. A more systematic procedure would be based on the empirical observation that optimum codes nearly always have efficiency quite close to one. Thus, one can compute

$$\mathcal{C}_{\min} = \sum_{i=1}^{n} p_i \log_t p_i,$$

and set $\mathcal{C}_0$ equal to $\mathcal{C}_{\min}/E$, where E is, say, 0.99. I a code having cost less than $\mathcal{C}_0$ is not found, a smalle value of E must be tried.

The success of the techniques for eliminating variable is best evidenced by the results of the integer programm ming calculations that have been carried out, using a experimental IBM 704 program. In these calculations, th constraint (31b) was used, rather than (31a), so that th b_j did not appear explicitly as variables. Also, if l_i ha the possible values $j_1, j_2, \cdots, j_s$, y_{ij_1} was represente as $1 - y_{ij_2} - y_{ij_3} \cdots - y_{ij_s}$. The results are given i Table I.

For codes 3 and 4, the alphabet encoded was the Roma alphabet plus "space," using the probabilities given b Brillouin[5] in 1956. Table II gives the costs of the cod

[5] Brillouin, [2], p. 52.

TABLE I

Summary of Computational Results

Code Number	1	2	3	4
Costs of Code Letters	1,2	1,2	1,2	2,3,3
n	10	12	27	27
Numbers of Variables	10	9	14	61
$\mathcal{C}_{\min} = \sum_{i=1}^{n} p_i \log_t p_i$	5.4164	4.8154	5.8270	6.6829
Cost of Solution	5.44	4.85	5.8599	6.7324
Efficiency of Solution	0.9957	0.9929	0.9944	0.9926
Computer Time Used	1 Minute	1 Minute	1 Minute	5 Minutes

TABLE II

Two Optimum Codes for the Roman Alphabet Plus "Space"

		CODE 3		CODE 4	
Letter	Probability of Letter	Cost of Code Word	Code Word	Cost of Code Word	Code Word
Space	0.200	3	1 1 1	4	2 2
E	0.105	5	2 1 1 1	5	3′ 2
T	0.072	6	1 1 2 2	6	3 3′
O	0.0654	6	1 2 1 2	6	3′ 3
A	0.063	6	2 1 1 2	6	3′ 3′
N	0.059	6	1 2 2 1	7	2 3 2
I	0.055	6	2 1 2 1	7	2 3′ 2
R	0.054	6	2 2 1 1	7	3 2 2
S	0.052	6	1 1 2 1 1	8	2 3′ 3′
H	0.047	6	1 2 1 1 1	8	3 2 3
D	0.035	7	1 2 1 1 2	8	3 2 3′
L	0.029	7	2 2 2 1	8	3 3 2
C	0.023	8	1 2 2 2 1	9	3 3 3
F	0.0225	8	2 1 2 2 1	9	3 3 3′
U	0.0225	8	2 2 1 2 1	10	2 3 3 2
M	0.021	8	1 1 2 1 2 1	10	2 3 3′ 2
P	0.0175	9	2 1 2 2 2	10	2 3′ 3 2
W	0.012	9	2 2 1 2 2	11	2 3 3 3′
Y	0.012	9	1 1 2 1 2 2	11	2 3 3′ 3
G	0.011	9	2 2 2 2 1	11	2 3 3′ 3′
B	0.0105	10	1 2 2 2 2 1	11	2 3′ 3 3
V	0.008	11	1 2 2 2 2 2	11	2 3′ 3 3′
K	0.003	12	2 2 2 2 2 2	14	2 3 3 3 3
X	0.002	13	2 2 2 2 2 1 1 1	14	2 3 3 3 3′
J	0.001	14	2 2 2 2 2 1 1 2	15	2 3 3 3 2 2
Q	0.001	14	2 2 2 2 2 1 2 1	16	2 3 3 3 2 3
Z	0.001	15	2 2 2 2 2 1 2 2	16	2 3 3 3 2 3′

ds in these codes, and gives specific prefix codes zing these costs. In code 3, the code letters are ted by 1 and 2; in code 4, they are denoted by 2, 3, 3′. Code 3 may be compared with one of cost 5.8629, ined by Marcus [14] in 1957 using a partially system- procedure.

V. Further Problems

ections II and III give several alternate algebraic racterizations of prefix codes. In contrast, an algebraic racterization of the larger class consisting of all quely decipherable codes has not been achieved, al- ugh Sardinas and Patterson [17] in 1953 have given nite procedure for deciding whether a code belongs this class. We conjecture that, for every uniquely ipherable code, there exists a prefix code having the e USF; Marcus [14] has made a somewhat similar jecture.

Turning to computational considerations, we note that, en the costs of code letters are high, and the channel acity therefore low, an inordinately large number of iables is required in the algorithm of Section IV. This y be remedied by introducing smaller integer costs ch are approximately proportional to the given costs, solving the resulting problem. This technique, how- r, yields only an approximate solution. Special tech- ues have been developed for obtaining strictly optimum es in this case, but their investigation has not yet been npleted.

Several problems of interest result when further con- aints are imposed on the coding problem. For example, practical communication systems, it may be desirable impose an upper bound on the costs of all code words. e coding problem with this constraint can be solved using the algorithm of Section IV; in fact, the externally posed upper bound simplifies the process by determining parameter m in advance. Huffman's method of con- ucting optimum codes, which is applicable when all le letters have equal cost, does not seem readily adapt- e to this constrained problem. Thus, even in the case equal costs, the algebraic approach has the advantage of ded flexibility. Table III gives length distributions for encodings of a twenty-seven letter alphabet, with babilities given by Gilbert and Moore [6] in 1959 for inary channel with $c_1 = c_2 = 1$. The first is the optimum le without constraints, and has cost 4.1195; the second optimum subject to the constraint that, for all i, ≤ 7, and has cost 4.1490.

Structures isomorphic to those of prefix codes occur many problems of classification and identification, sing in such diverse fields as sorting, character recogni- n, and medical diagnosis. Such problems have been sidered by Schützenberger [18] in 1954 and Moore[6].

[6] E. F. Moore, private communication; July, 1959.

As an example, suppose we are given the probabilities of occurrence of a set of objects, and a set of tests, having known outcomes for each object, which may be applied to determine the identity of an unknown object. Such a situation is specified in Table IV, in which each test has the two possible outcomes, zero and one.

TABLE III

Result of Coding with Constraints

Letter	Probability	Length in Unconstrained Code	Length in Constrained Code
Space	0.1859	3	3
E	0.1031	3	3
T	0.0796	4	4
A	0.0642	4	4
O	0.0632	4	4
I	0.0575	4	4
N	0.0574	4	4
S	0.0514	4	4
R	0.0484	4	4
H	0.0467	4	4
L	0.0321	5	5
D	0.0317	5	5
U	0.0228	5	6
C	0.0218	5	6
F	0.0208	6	6
M	0.0198	6	6
W	0.0175	6	6
Y	0.0164	6	6
G	0.0152	6	6
P	0.0152	6	6
B	0.0127	6	6
V	0.0083	7	7
K	0.0049	8	7
X	0.0013	10	7
J	0.0008	10	7
Q	0.0008	10	7
Z	0.0005	10	7

TABLE IV

Object X_i	Probability p_i	Tests a	b	c	d	e	f
X_1	0.23	1	1	1	0	0	0
X_2	0.20	1	1	0	0	1	0
X_3	0.16	0	1	0	0	1	1
X_4	0.11	1	0	1	0	1	0
X_5	0.08	1	0	0	0	1	1
X_6	0.07	0	0	1	1	1	1
X_7	0.05	1	0	0	1	1	1
X_8	0.04	1	1	1	0	1	1
X_9	0.03	1	0	1	1	0	1
X_{10}	0.03	1	1	1	1	0	1

An identification procedure for this example is shown in Fig. 3. At each node of the tree a test is performed, and one branch or the other is followed, depending on the outcome of the test. Unique identification is achieved when a terminal node is reached. The expected number of tests required for an identification is $\sum p_i l_i$, where l_i is the length of the path leading to the terminal node associated with X_i.

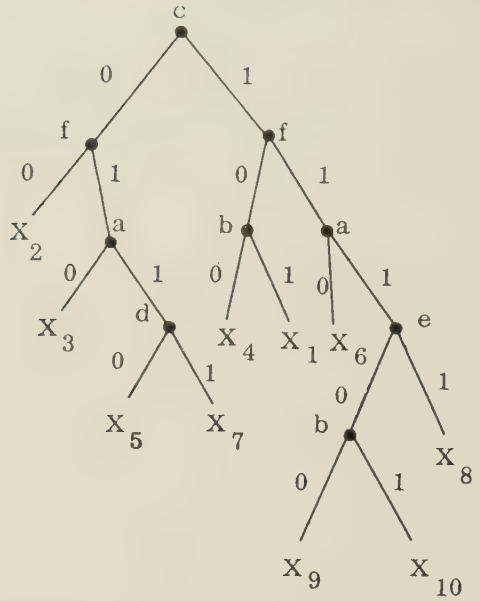


Fig. 3—Tree for an identification procedure

Suppose that the optimum length distribution for the given alphabet $\{X_i\}$ has been found. Then it is not difficult to determine whether this length distribution is realizable by the given set of tests. If it is not realizable, the 2nd best, 3rd best, $\cdots$, kth best, $\cdots$ distributions may be considered in turn. This is done using integer programming by inserting at each stage a constraint which rules out the last distribution obtained, and no other. The incremental cost of obtaining each such solution, once the first has been obtained, is not great. Of course, if the given set of tests is not highly effective, many distributions will have to be inspected, and the method will not be feasible. It is hoped that further investigation along these lines will yield improved methods for treating classification problems.

Bibliography

[1] N. M. Blachman, "Minimum-cost encoding of information IRE Trans. on Information Theory, vol. IT-3, pp. 139–14 March, 1954.
[2] L. Brillouin, "Science and Information Theory," Academ Press, Inc., New York, N. Y.; 1956.
[3] G. B. Dantzig, draft of "Discrete variable extremum problems in "Linear Programming and Extensions," ch. 26, to be pu lished.
[4] L. E. Dickson, "History of the Theory of Numbers," Carneg Institute of Washington, Washington, D. C., vol. 1; 191
[5] R. M. Fano, Res. Lab., for Electronics, Mass. Inst. Tech Cambridge, Tech. Rept. no. 65; 1949.
[6] E. N. Gilbert and E. F. Moore, "Variable length binar encodings," *Bell Sys. Tech. J.*, vol. 38, pp. 933–967; Jul 1959.
[7] R. E. Gomory, "Outline of an algorithm for integer solutio to linear programs," *Bull. Am. Math. Soc.*, vol. 64, pp. 275–27 September, 1958.
[8] R. E. Gomory, "An algorithm for integer solutions to linea programs," Princeton-IBM Math. Res. Project, Tech. Rep Rept. no. 1; 1958.
[9] R. E. Gomory, "All-Integer Integer Programming Algorithm IBM Res. Center, Yorktown Heights, N. Y., Res. Rep RC-189; 1960.
[10] D. A. Huffman, "A method for the construction of minimu redundancy codes," Proc. IRE, vol. 40, pp. 1098–110 September, 1952.
[11] L. G. Kraft, 'A Device for Quantizing, Grouping, and Codin Amplitude Modulated Pulses," M. S. Thesis in electric engineering, Mass. Inst. Tech., Cambridge; 1949.
[12] J. L. Lagrange, *Nouv. Mem. Ac. Berlin*, pp. 183–272; 1775
[13] B. Mandelbrot, "On recurrent noise limiting coding," *Symp on Information Networks*, New York, N. Y., pp. 205–221 April, 1954.
[14] R. S. Marcus, "Discrete Noiseless Coding," M. S. thesis i electrical engineering, Mass. Inst. Tech., Cambridge; 1957
[15] M. Marden, "The Geometry of the Zeros of a Polynomia in a Complex Variable," Am. Math. Soc., New York, N. Y 1949.
[16] B. McMillan, "Two inequalities implied by unique decipher ability," IRE Trans. on Information Theory, vol. IT-2 pp. 115–116; December, 1956.
[17] A. A. Sardinas and G. W. Patterson, "A necessary and sufficien condition for the unique decomposition of coded messages, 1953 IRE Convention Record, pt. 8, pp. 104–108.
[18] M. P. Schützenberger, "Contribution aux Applications Sta tistiques de la Theorie de L'Information," Université de Paris Paris, France, vol. 3, Fascicules 1–2; 1954.
[19] C. E. Shannon, 'A mathematical theory of communication,' *Bell Sys. Tech. J.*, vol. 27, pp. 379–423, 623–656; July/October 1948.

Note on Signal-to-Noise Ratio in Band-Pass Limiters*

CHARLES R. CAHN†, MEMBER, IRE

mmary—A simplified analysis is presented to explain physically change of signal-to-interference ratio which occurs in a band- limiter. The analysis utilizes the concept of sideband resolution symmetric and anti-symmetric parts and considers only the ptotic case where the signal-to-interference ratio is small in parison with unity. Wide-band correlation-detection systems liscussed, as well as ordinary band-pass systems.

ie important conclusion is reached that the degradation is ly dependent on the statistics of the interference amplitude uations. However, when the signal is weak compared to the ference, the maximum possible degradation is 6 db and occurs onstant-amplitude interference.

egradation with noise interference in a wide-band correlation- ction system has been obtained for arbitrary signal and noise lwidths. It is found that the degradation ranges between 0.6 db 1.0 db, the latter figure being for the case where the signal lwidth is greater than approximately three times the noise lwidth.

Introduction

THE EFFECT of an ideal band-pass limiter on signal-to-noise ratio has been rigorously analyzed[1] for the case of a sine wave in Gaussian noise inter- nce. The results indicate that the input signal-to-noise o is degraded only by a numerical factor close to unity, maximum degradation being $4/\pi$ (1.0 db) for signal- oise ratios much less than unity (0 db). This basic lt has been extended[2] to allow calculation of the effect he band-pass limiter on signal detectability, assuming eak signal-to-noise ratio at the limiter input and ssian noise interference. The results show that the ter produces a very small loss in signal detectability; example, the factor is only 1.16 (0.6 db) for a wide- d rectangular noise spectrum.

he analyses referred to in the above paragraph do admit an easily grasped physical picture which shows rly why the desired signal in the presence of strong rference is not highly suppressed by the nonlinear on of the limiter. Furthermore, the analyses are valid for Gaussian noise interference and do not present ected performance for arbitrary interference statistics. rther limitation is the restriction that the interference dwidth be much wider than the signal bandwidth. is the purpose of this paper to present a simple lysis which, while applicable only for the case of a al much weaker than the interference, does provide a simple physical picture and does not have the limitations mentioned above. The analysis will lead to a degradation factor which relates the output signal-to-interference ratio to the input ratio. However, as in the analyses mentioned,[1,2] only long-term averages are considered. It is, of course, recognized that short-term fluctuations can be significant in many practical situations where the application of a limiter might be considered. In addition, the interference is assumed noncoherent with the signal. A special analysis is required to treat coherent interference, which can greatly suppress the desired signal in certain cases.

The use of signal-to-interference ratio as an indication of system performance is common, in practice, for simplicity and is adopted here for this reason. It is, of course, true that the problem of signal detection has been studied extensively, and a more accurate and general statistical detection theory has been evolved to replace the simple criterion of signal-to-interference ratio.[3]

Case of Two Sine Waves

The case which serves as the basis for simple analysis of more general inputs is that of two sine waves. For this case, the limiter input may be expressed as $A_1[\cos \omega_1 t + a \cos \omega_2 t]$, where a denotes the amplitude ratio and is less than unity. The output of an ideal band-pass limiter, which removes the amplitude modulation without distorting the phase modulation,[4] is obtained by dividing the input by the instantaneous envelope, yielding

$$\frac{A_1[\cos \omega_1 t + a \cos \omega_2 t]}{A_1\sqrt{1 + a^2 + 2a \cos (\omega_1 - \omega_2)t}} = \cos \omega_1 t + \frac{a}{2} \cos \omega_2 t - \frac{a}{2} \cos (2\omega_1 - \omega_2)t + \frac{\text{terms proportional to}}{\text{higher powers of } a.} \quad (1)$$

It is seen from (1) that the limiter suppresses the weaker signal, relative to the stronger signal, by an amplitude factor of 2 (6db) and produces cross-modulation components, only one of which is of significant amplitude. In addition, the phase of each of the two signals is not affected by the limiter.

The above result, obtained by a series expansion method, is more easily established by considering the weaker signal as a sideband of the stronger signal and using the concept of symmetric and anti-symmetric sidebands.[5] It

Received by the PGIT, May 3, 1960; revised manuscript ved, July 19, 1960.

Bissett-Berman Corp., Los Angeles, Calif. Formerly with e Tech. Labs., Inc., Los Angeles, Calif.

W. B. Davenport, "Signal-to-noise ratios in band-pass limiters," *ppl. Phys.*, vol. 24, pp. 720–727; June, 1953.

R. Manasse, R. Price, and R. M. Lerner, "Loss of signal ctability in band-pass limiters," IRE TRANS. ON INFORMATION ORY, vol. IT-4, pp. 34–38; March, 1958.

[3] D. Van Meter and D. Middleton, "Modern statistical approaches to reception in communication theory," IRE TRANS. ON INFORMATION THEORY, no. PGIT-4, pp. 119–145; September, 1954.

[4] W. B. Davenport and W. L. Root, "Introduction to the Theory of Random Signals and Noise," McGraw-Hill Book Co., Inc., New York, N. Y., p. 288; 1958.

[5] S. Goldman, "Frequency Analysis, Modulation and Noise," McGraw-Hill Book Co., Inc., New York, N. Y., pp. 167–181; 1948.

is easily shown that the symmetric sidebands produce amplitude modulation only, and to a first approximation for weak sidebands, the anti-symmetric sidebands produce phase modulation only. Since the ideal band-pass limiter suppresses the amplitude modulation, only the carrier and the anti-symmetric sidebands are retained in the limiter output. The first-order terms on the right side of (1) are observed to be exactly these components. This approach may be generalized to include a multiplicity of frequency components about a single strong carrier. Each component is found to be independently affected by the limiter to the first approximation and, accordingly, is suppressed by an amplitude factor of 2 and is unchanged in phase.

Since the relative amplitudes and phases of the various frequency components of the desired signal are unchanged despite strong sine wave interference, the significant observation is made that the limiter output contains an undistorted replica of the desired signal. In fact, this conclusion applies even if the sine wave interference is phase modulated, and explains physically why the desired signal is not highly suppressed by the strong interference.

Case of a Sine Wave and Strong Gaussian Noise Interference

If the interference is Gaussian noise and strong compared to the sine wave, it may be considered as a modulated carrier. The input to the limiter then may be expressed as

$$\text{input} = A_n \cos(\omega_1 t + \theta_n) + \sqrt{2S}\cos\omega_2 t, \tag{2}$$

in which the phase θ_n of the noise interference is random and the amplitude A_n has a Rayleigh distribution. For simplicity, ω_1 and ω_2 are assumed unequal, although a more complicated argument can be used if $\omega_1 = \omega_2$ to yield the same result. For a fixed noise amplitude A_n, the sine wave, being a weak sideband, gives rise to two output sine waves of equal amplitude, one of which is at the frequency ω_2 and in phase with the input sine wave. The other, at the frequency $2\omega_1 - \omega_2$, has the random phase of the noise. Although both output sine waves have the amplitude $\sqrt{S}/\sqrt{2}\,A_n$, there will be no average output at the frequency $2\omega_1 - \omega_2$, because of the random phasing.

The average amplitude (or steady component) of the sine wave output at the frequency ω_2 is obtained by averaging over all possible noise amplitudes, as follows:

$$\text{Average sine-wave amplitude} = \sqrt{\frac{S}{2}}\,\overline{A_n^{-1}}$$

$$= \sqrt{\frac{S}{2}}\int_0^\infty \frac{1}{A_n}\frac{A_n}{N}e^{-A_n^2/2N}\,dA_n = \sqrt{\frac{\pi}{4}\frac{S}{N}} \tag{3}$$

where the bar denotes an ensemble average and N is the average power of the interference. The output noise essentially has a power of $\frac{1}{2}$, since the limiter output is a phase-modulated sinusoid of unit amplitude and the desired signal component is much smaller than the noise component. Thus, the output signal-to-interference ratio

$$\left(\frac{S}{N}\right)_{\text{out}} = \frac{\left(\sqrt{\frac{\pi}{4}\frac{S}{N}}\right)^2/2}{1/2} = \frac{\pi}{4}\left(\frac{S}{N}\right)_{\text{in}},$$

which is identical with the result obtained by Davenpor by a rigorous treatment of this case.

It should be noted that with Gaussian noise inter ference, the average sine wave amplitude in the limite output can be evaluated in closed form for an arbitrar signal-to-interference ratio directly from (2) divided b the instantaneous envelope.[6,7] Since the total outpu power from the limiter is always $\frac{1}{2}$, the output signal-t interference ratio can also be expressed in closed form However, a similar closed-form result for non-Gaussia interference does not appear to exist, in general.

Case of a Sine Wave and Strong Non-Gaussian Interference

The calculation made for Gaussian noise (Rayleig distribution of amplitude) can be generalized to includ interference with an arbitrary amplitude distribution, o the assumption that the output interference power i much larger than the output signal power. From (3) an the fact that the output interference power is essentially $\frac{1}{2}$ the output signal-to-interference ratio is found to b

$$\left(\frac{S}{N}\right)_{\text{out}} = \frac{S}{2}\,\overline{(A_n^{-1})^2}. \tag{5}$$

On the other hand, the input signal-to-noise ratio i

$$\left(\frac{S}{N}\right)_{\text{in}} = \frac{2S}{\overline{A_n^2}}. \tag{6}$$

Therefore, the degradation in signal-to-interference rati due to the limiter is given by the factor

$$\Lambda = \frac{(S/N)_{\text{in}}}{(S/N)_{\text{out}}} = \frac{4}{\overline{A_n^2}\,\overline{(A_n^{-1})^2}}. \tag{7}$$

As an example, (7) may be used to derive a closed-form expression for the degradation when the interference is combination of a steady component and a Gaussian nois component. The amplitude of this combination has th probability density function

$$p(A_n) = A_n e^{-(A_n^2+2\gamma)/2} I_0(\sqrt{2\gamma}A_n) \tag{8}$$

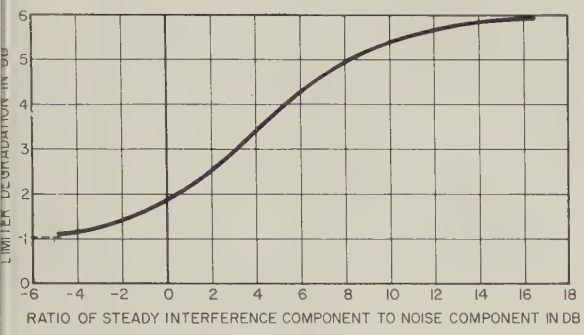
where γ is the power ratio of the steady component an the noise component. If the averages in (7) are evaluated the result is obtained as follows:

$$1/\Lambda = \frac{\pi}{4}(\gamma+1)[e^{-\gamma/2}I_0(\gamma/2)]^2, \tag{9}$$

which is graphed in Fig. 1.

[6] N. M. Blackman, "The output signal-to-noise ratio of a powe law device," *J. Appl. Phys.*, vol. 24, pp. 783–785; June, 195

[7] I. S. Reed, "An Approximation to the Output Signal-to-nois Ratio of a Signal, Passed through a Band-Pass Limiter, Followe by a Narrow-Band Filter," Lincoln Lab., Mass. Inst. Tech., Lex ington, Group Rept. 47.17; June 2, 1958.



1—Degradation for a combination of steady and Gaussian interference.

he graph in Fig. 1 shows how the degradation due ne limiter increases to a limiting value of 6 db as the uation in the interference amplitude decreases. The st possible degradation (6 db) actually occurs for rference with a constant amplitude, as may be verified pplication of the Schwarz inequality.[8] The proof con- of verifying the successive inequalities,

$$\overline{A_n^2}(\overline{A_n^{-1}})^2 \geq (\overline{A_n}\ \overline{A_n^{-1}})^2 \geq 1. \tag{10}$$

inequalities become equalities only if

$$\overline{A_n^2} = (\overline{A_n})^2, \tag{11}$$

ch means that the variance of the probability distri- on of amplitude is zero. This occurs only if the litude is constant. On the other hand, there is no r limit; that is, probability distributions with the ropriate behavior at $A_n = 0$ render (7) arbitrarily ll.[9] This nonmeaningful result arises from the ap- ximation that the interference should always be strong pared to the sine wave. Furthermore, the signal-to- rference ratio is not an accurate criterion of system ormance in such a case.

nce the interference is assumed to be noncoherent, average output exists at the frequency $2\omega_1 - \omega_2$, this term is not considered further in the discussion.

Wide-Band Correlation Systems

communication system can utilize a wide-band signal a complex waveform, the objective being to trade dwidth for interference reduction. Such a system may ze correlation detection as an effective narrow-band ption technique.[10,11] While correlation detection with atched reference is an optimal procedure for signals corrupted by additive white Gaussian noise,[12] it still often is used in practice when interference of a more general nature is present. For practical reasons related to gain control, a limiter can be incorporated in the receiver, as indicated in Fig. 2, which illustrates the basic coherent type of correlation processing. It is desired to calculate the degradation introduced by the limiter, determined by comparing the output signal-to-interference ratios obtained with and without the limiter. The restriction is made that the interference is much stronger than the signal, which is the case in practical situations where wide-band signals are used for interference reduction.

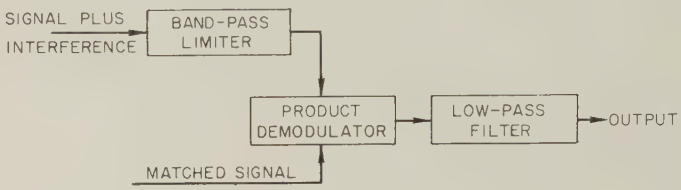


Fig. 2—Coherent correlation system.

Since the assumptions that the interference is strong and the signal is wide-band usually ensure that the output interference from the narrow-band low-pass filter is essentially Gaussian even when the limiter is used, the calculated signal-to-interference ratio degradation can be interpreted directly as a performance degradation.

The input to the receiver may be expressed as $s(t) + n(t)$, where $s(t)$ is the desired signal waveform and $n(t)$ is the interference waveform. The calculation of output signal-to-interference ratio will be performed first in the absence of limiting. Then the product demodulator output is $s^2(t) + s(t)n(t)$, the first term of which is the output signal and the second is the output interference. Thus, the average or dc output is

$$\text{dc output} = \overline{s^2(t)} = S, \tag{12}$$

using S for the average power of the desired signal $s(t)$. The output interference term $s(t)n(t)$ is the product of two independent waveforms, so that its autocorrelation function is the product of the individual autocorrelation functions. The power spectral density of the output interference can therefore be obtained by convolution of the respective power spectral densities, $S(f)$ and $N(f)$, of the input signal and interference. Since the output interference spectrum is essentially constant over the significant passband of the low-pass filter, only the zero-frequency value $N_{out}(0)$ is needed and is given by

$$N_{out}(0) = \int_0^\infty S(f)N(f)\,df, \tag{13}$$

in which one-sided spectral densities are utilized.

Birkhoff and MacLane, "A survey of Modern Algebra," The Millan Co., New York, N. Y., p. 183; 1948. The proof using chwarz inequality was initially developed by Dr. R. E. Graves, e Tech. Labs.

For example, distributions with a behavior at zero amplitude e form A_n^α, where $\alpha \leq 0$, cause the second average in (7) divergent.

P. E. Green, Jr., ' The output signal-to-noise ratio of correla- detectors," IRE Trans. on Information Theory, vol. IT-3, 0-18; March, 1957.

R. Price and P. E. Green, Jr., A communication technique ultipath channels," Proc. IRE, vol. 46, pp. 555–570; March,

[12] P. M. Woodward, "Probability and information Theory with Applications to Radar," McGraw-Hill Book Co., Inc., New York, N. Y.; 1953.

The output interference power is equal to the product of $N_{out}(0)$ and the noise bandwidth b of the low-pass filter.[13] Thus, the output signal-to-interference ratio is given by

$$\left(\frac{S}{N}\right)_{out} = \frac{S^2}{b\int_0^\infty S(f)N(f)\,df}, \tag{14}$$

which is a special case of (9) of Green's article.[10] In addition, fluctuations of the signal term $s^2(t)$ in the product demodulator output will also be present in the output of the low-pass filter. This fluctuation has been called self-noise, despite the fact that it is completely predictable from the wave-form $s(t)$.[10] However, when the interference is strong, the self noise is negligible, and for this reason has not been included in (14).

When the interference spectrum is uniform, $N(f) = N_0$, (14) may be shown to reduce to the well-known formula for the peak signal-to-noise ratio from a matched filter.[14] To demonstrate this, the low-pass filter is assumed to be an ideal integrator with a rectangular impulse response of duration T.[15] The noise bandwidth of this filter is easily calculated to be $1/2T$, so that (14) becomes

$$\left(\frac{S}{N}\right)_{out} = \frac{2TS^2}{N_0\int_0^\infty S(f)\,df} = \frac{2TS}{N_0} = \frac{2E}{N_0} \tag{15}$$

where $E = TS$ is the energy of the signal over the time interval T.

When limiting is performed on the input, the output signal-to-interference ratio becomes dependent on the amplitude distribution of the interference. Since the interference is strong, it is convenient to express the desired signal as a superposition of various frequency components and the interference as a modulated carrier with amplitude A_n, as in (2). Comparison of (2) with the first part of (3) shows that the limiter reduces the amplitude of each frequency component of the desired signal by the factor $\overline{A_n^{-1}}/2$, so that the useful dc output amplitude is also reduced by this factor. The output interference power has only a negligible contribution from the desired signal and may be evaluated from knowledge of the spectral density of the interference at the limiter output by an integral similar to (13). This distorted spectral density will be denoted by $N_L(f)$ and has a total power content of $\frac{1}{2}$. The degradation in output signal-to-interference ratio produced by the limiter then may be expressed as

$$\Lambda = \frac{(S/N)_{\text{no limiter}}}{(S/N)_{\text{limiter}}} = \frac{4\int_0^\infty S(f)N_L(f)\,df}{(\overline{A_n^{-1}})^2\int_0^\infty S(f)N(f)\,df}. \tag{16}$$

It should be emphasized at this point that the degradation Λ defined by (16) is not, for general interference spectra, a degradation from an optimum detection process, but only an expression of the effect of a limiter in a system using correlation detection with a matched reference.

If $S(f)$ is essentially constant with the value $S(f_0)$ over the significant portions of $N(f)$ and $N_L(f)$ (narrow-band interference), (16) may be approximated as

$$\Lambda = \frac{4S(f_0)\int_0^\infty N_L(f)\,df}{(\overline{A_n^{-1}})^2 S(f_0)\int_0^\infty N(f)\,df} = \frac{4}{(\overline{A_n^{-1}})^2\,\overline{A_n^2}}, \tag{17}$$

since the integral in the numerator is simply $\frac{1}{2}$, and the integral in the denominator is $\overline{A_n^2}/2$. Eq. (17) is identical with (7). Hence, when a limiter is employed, the maximum degradation with narrow-band interference is 6 db, following the same argument used in connection with (7) and occurs with constant-amplitude interference. Actually this conclusion is independent of interference bandwidth when the interference amplitude is constant, since $N_L(f)$ is proportional to $N(f)$ (no distortion), so that (16) reduces to a value of 4.

The only other example which will be treated in detail is the case where the interference is Gaussian noise and both noise and signal have rectangular spectra with the same center frequency f_0. The bandwidths are denoted by B_N and B_S, and the powers by N and S, respectively. The spectral density $N_L(f)$ of the interference at the limiter output is indicated in Fig. 3, using a result of Price.[16] The average of A_n^{-1} in the denominator of (16) may be evaluated as shown in (3). The value of the integral in the denominator is either SN/B_S or SN/B_N depending on whether $B_S > B_N$ or $B_S < B_N$. It is then found that the expression for the degradation due to the limiter is either

$$\Lambda = \frac{8}{\pi}\int_{f_0-B_S/2}^{f_0+B_S/2} N_L(f)\,df \qquad (B_S > B_N) \tag{18}$$

or

$$\Lambda = \frac{8B_N}{\pi B_S}\int_{f_0-B_S/2}^{f_0+B_S/2} N_L(f)\,df \qquad (B_S < B_N). \tag{19}$$

Note that the integral in either (18) or (19) specifies the total output interference power contained within the bandwidth B_S of the desired signal.

[13] J. L. Lawson and G. E. Uhlenbeck, "Threshold Signals," McGraw-Hill Book Co., Inc., New York, N. Y., p. 176; 1950.

[14] G. L. Turin, "An introduction to matched filters," IRE TRANS. ON INFORMATION THEORY, vol. IT-6, pp. 311–329; June, 1960.

[15] David Middleton, "An Introduction to Statistical Communication Theory," McGraw-Hill Book Co., Inc., New York, N. Y., p. 683; 1960.

[16] R. Price, "A note on the envelope and phase-modulated components of narrow-band Gaussian noise," IRE TRANS. ON INFORMATION THEORY, vol. IT-1, pp. 9–15; September, 1955.

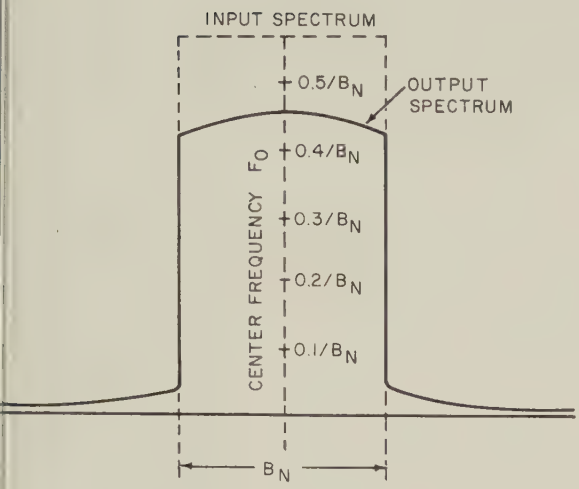


Fig. 3—Output spectrum of band-pass limiter.

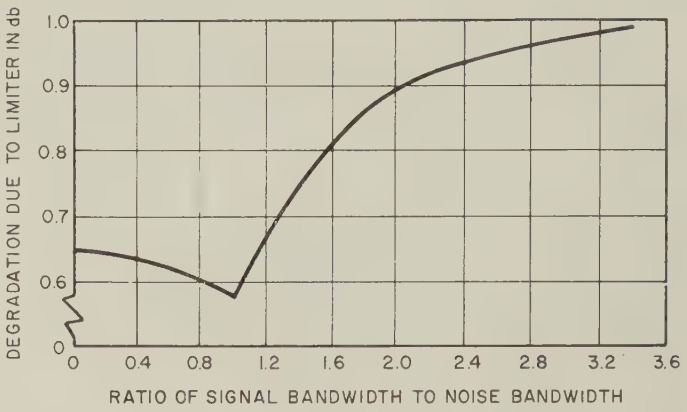


Fig. 4—Degradation in correlation system with Gaussian Interference.

wo extremes may be considered. First, let the signal dwidth be much larger than the noise bandwidth, so t the "tails" of $N_L(f)$ are included in the integral of). The integral then is simply $\frac{1}{2}$, and $\Lambda = 4/\pi$ (1.0 db). ond, let the noise bandwidth be much larger than the al bandwidth, so that (19) becomes

$$\Lambda = \frac{8}{\pi}\frac{B_N}{B_S} B_S(0.456/B_N) = 1.16 \qquad (20)$$

0.6 db, the result obtained by Manasse, Price, and ner,[2] when the signal is assumed at the center of the se spectrum. The transition between the two extremes hown in Fig. 4. It is interesting to note that the least radation occurs when the signal and noise have the e bandwidth, although from a practical point of view variation with bandwidth is very small.

Conclusions

he simplified analysis of the effect of an ideal band- limiter on signal-to-interference ratio has led to lts in agreement with those obtained by more rigorous techniques. In addition, the method obtains useful answers for non-Gaussian interference and may be applied to both ordinary band-pass systems and wide-band correlation-detection systems.

The most important conclusion is that the degradation in signal-to-interference ratio due to the presence of a limiter is quite dependent on the statistics of the interference amplitude. However, when the signal-to-interference ratio is low, the worst degradation (6 db) results for constant amplitude interference in both ordinary band-pass and wide-band correlation-detection systems. Amplitude fluctuations will reduce the degradation considerably; for example, the degradation with noise-like interference is about one db. The degradation (in db) may be expected to be negative for highly fluctuating interference (for example, impulsive interference). This indicates an important practical reason for incorporating a limiter into a communication system.

Finally, the result is obtained that with Gaussian noise interference the degradation due to a limiter in a wide-band correlation-detection system does not vary significantly with interference bandwidth. Previous analyses have been restricted to the case of wide-band interference only.

Recognition of Membership in Classes*

GEORGE S. SEBESTYEN†, MEMBER, IRE

Summary—This paper presents an approach to the general problem of recognition of membership in classes which are known only from a set of their examples. A geometrical approach is taken where membership in classes is regarded measurable by metrics with which a set of points, representing different members of the same class, may be brought "close" to one another. For the case where classes are Gaussian processes, the method described herein and that of decision theory are found to agree. A practical application of the method to the automatically "learned" recognition of spoken numerals is described.

Introduction

As the advances of modern science and technology furnish the solutions to problems of increasing complexity, a feeling of confidence is created in the realizability of mathematical models or machines which can perform *any* task as long as a specified set of instructions can be given stating how the task is to be performed. There are, however, problems of long-standing interest which have hitherto eluded solution, partly because the problems have not been clearly defined, and partly because no specific instructions could be given stating how a solution should be reached. Recognition of a spoken word independent of the speaker who utters it, recognition of a speaker regardless of the spoken text, the problem of threat evaluation, that of making a medical diagnosis, and that of recognizing a person from his handwriting are only a few of the problems which so far remained largely unsolved for the above mentioned reasons.

In all of the problems of pattern recognition, however different they may seem, there is a common bond that unites them and permits their solution with identical methods. The common bond is that the solution of these problems requires the ability to recognize membership in classes, and, more important, it requires the automatic establishment of *how* to measure membership in each class. In word recognition, the class is a specific word of interest and members of the class are different utterances of the word by different speakers. If membership in the class could be recognized, then the *word*, independent of the speaker, could be identified. Similarly, "speech by a given speaker" or "handwriting by a given person" are classes in which membership must be recognized in solving the problems listed above. In a similar manner, the rendering of a medical diagnosis is the recognition of the patient as a member of the class of individuals having a particular-disease, while threat evaluation (say the decision of "attack" or "no attack") consists of the recognitio of the present situation as a member of the class c situations which constitute threat.

The purpose of this paper is to present a heuristi approach to the recognition of membership in classe known only from a given set of their examples. Firs the fundamental ideas and underlying assumptions o which the theoretical approach is based are discussed A mathematical embodiment of these ideas is then de veloped. For a special selection of classes, agreement wit a decision-theoretical approach is demonstrated. In furthe support of this approach, results obtained in its successfu experimental verification are described.

Fundamental Ideas and Assumptions

The desired objective is to find automatic methods fo learning how to measure membership in classes that ar known only through a set of their examples. It will b assumed in the following that representative examples o the classes to be recognized are given and from thes examples class definitions are to be constructed whic can be used to classify correctly each new occurrence a a member of the class to which it actually belongs. W will consider a general event—a member of any of th classes—represented by a point or vector in an N-dimen sional observation space which serves as a model of th physical world. Each dimension expresses a property o the event, a type of observation that can be made about it The entire signal which represents all the informatio available about the event is a vector $v = (v_1, v_2, \cdots, v_N)$ the coordinates of which have numerical values whic correspond to the amount of each property which th event contains. In this representation, a set of event that belong to the same class correspond to an ensembl of points scattered within some region of the observatio space. A set of sample points from a given class migh be expected to "cluster" in the N-dimensional space i the sense that distances between members of the sam class are smaller, on the average, than those betwee points that belong to different classes. Unfortunately this state of affairs cannot generally be expected to exis Therefore the concept which plays a central role in th theory which will be described is the notion that point in the observation space which represent a set of non identical events belonging to a common class must b close to one another as measured by *some* as yet unknow method of measuring distance, since the points represen events which are close to one another in the sense tha they are members of the same class. Mathematicall speaking, the fundamental notion underlying the theor is that similarity (closeness in the sense of belonging t

* Received by the PGIT, August 17, 1960. The work reported in this article was undertaken as part of the D.Sc. dissertation work at Mass. Inst. Tech., Cambridge, Mass., and was continued under Contract AF 30(602)-2112 at Melpar, Inc., Boston, Mass.

† Litton Systems, Inc., Advanced Dev. Lab., Waltham, Mass.; formerly with Melpar, Inc., Boston.

same class or category) is expressible by a metric method of measuring distance), by which points resenting examples of the class we wish to recognize found to lie close to each other.

To give credence to this conjecture, consider what is ant by the abstract concept of a class. According to of the possible definitions, a class is a collection of ngs which have some common properties. By a modifi- ion of this thought, a class could be characterized by common properties of its members. A metric by which nts representing examples of a class are close to each er must therefore operate chiefly on the common perties of the examples and must ignore, to a large ent, those properties not present in each example. nsequently, if a metric were found which called examples he class "close," somehow it must exhibit their common perties.

To present this fundamental idea in a slightly different y, we can state that a transformation on the observa- n space which is capable of clustering the points repre- ting the examples of the class must operate primarily the common properties of the examples. A simple stration of this idea is shown in Fig. 1, where the

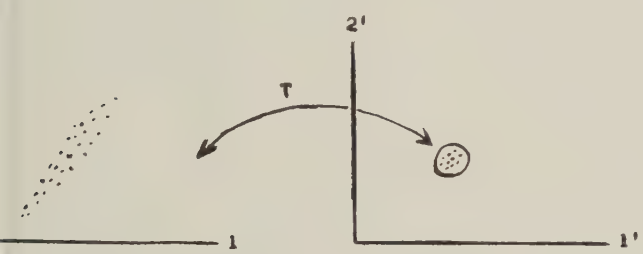


Fig. 1—Clustering by transformation.

emble of points is spread out in the observation space ly a two-dimensional space is shown for ease of illu- ation), but a transformation T of the space is capable clustering the points of the ensemble. In the above mple neither the signal's property represented by rdinate 1 nor that represented by coordinate 2 is ficient to describe the class, for the spread in each is ge over the ensemble of points. Some function of the coordinates, on the other hand, would exhibit the mmon property that the dispersion of the points about xed straight line is small. In this specific instance, of rse, correlation between the two coordinates would ibit this property; but in more general situations ple correlation will not suffice.

f the observation space were flexible (like a rubber et), the transformation T would express the manner which various portions of the space must be stretched compressed in order to bring the points together most sely.

Although thinking of transformations of the space is as general as thinking about exotic ways of measuring stance" in the original space, the former is a rigorously rect and easily visualized analogy for many important sses of metrics.

As any mathematical theory, the one which evolved from the preceding ideas is based on certain assumptions. The most basic assumption is that the N-dimensional space in which events exemplifying their respective classes are represented is a complete enough model of the physical world to contain information about the common properties which serve to characterize the classes. The significance of this assumption is appreciated if we consider, for example, that the observation or signal space contains all the information that a black and white television picture could present of the physical objects making up the set of events which constitute the examples of a class. No matter how ingenious the data processing schemes that we might evolve may be, objects belonging to the class "red things" could not be identified, because representation of the examples by black and white television simply does not contain color information. For any practical situation one must rely on engineering judgment and intuition to determine whether the model of the real world (the observation space) is complete enough. Fortunately, in most cases, this determination may be made with considerable confidence.

A second assumption states the class of transformations or the class of metrics within which we look for the "best." This assumption is equivalent to specifying the allowable methods of stretching or compressing the observation space within which we look for the best specific method of deforming the space. In effect, an assumption of this type specifies the type of network (such as active linear networks) to which the solution is restricted.

The third major assumption is hidden in the statement that we are able to specify when the solution is best. In practice, of course, we can frequently say what is considered a good solution even if we do not know which is the best. The criterion by which the quality of a metric or transformation is judged good is thus one of the basic assumptions.

The sufficiency of the examples as a representative set is also an assumption which needs to be considered. It is perhaps the most important assumption for, in practice, concurrence regarding the sufficiency of a set of examples is most difficult to obtain. Within the constraints of these assumptions the mathematical embodiment of the fundamental ideas will now be outlined.

Minimization of Mean-Square Distance

The task of learning how to measure membership in classes consists of partitioning the observation space into regions in a manner which depends optimally on the distribution of the known sample points in the space. In the above process, sets of examples of each of several classes are assumed given, and each sample point is labeled with the name of the class to which it is *a priori* known to belong. Analogous to the application of likelihood ratios in decision theory, a convenient way to partition the observation space into regions, one cor-

responding to each class, is to generate for each class a function which gives a quantitative measure of how "close" an arbitrary point of the space is to members of a specific class. In a sense, each function measures the similarity of an arbitrary point to a class, and partitioning the space is accomplished by assigning the point to that class to which it is most similar. We will arbitrarily choose the mean-square distance between a point x and known members of the class to serve as a quantitative measure of similarity S between x and the class. This definition is expressed by (1), where f_m is the mth known member of class F, $d(\ )$ is an as yet unspecified metric which expresses the sense in which members of F are closest to one another, and M is the number of given members of F;

$$S(x, \{f_m\}) = \frac{1}{M} \sum_{m=1}^{M} d^2(x, f_m). \tag{1}$$

The criterion for selecting the metric is that, if distance is measured by the optimum one from a class of metrics, then the mean-square distance between members of F (a measure of clustering) should be minimum. Note that the unknown in the minimization is the metric which is the error criterion in measuring similarity.

In the following derivation, we will carry out the minimization for a simple class of metrics obtained by considering the Euclidean distance measured on a linear transformation of the space, where the transformation is subject to a suitable constraint devised to assure a nontrivial solution. The Euclidean distance, after linear transformation of the space, is expressed by (2a), and a constraint which assures a nontrivial solution is given in (2b). The constraint prevents minimization by the general collapse of the space by fixing the product of the squared lengths of the sides of all N-dimensional parallelepipeds.[1] By Hadamard's theorem, this constraint also holds volume invariant under orthogonal transformations.

$$d(f, g) = \sqrt{\sum_{i=1}^{N} \left[\sum_{j=1}^{N} a_{ij}(f_j - g_j) \right]^2}; \tag{2a}$$

$$1 = \prod_{i=1}^{N} \left(\sum_{j=1}^{N} a_{ij}^2 \right) = \prod_{i=1}^{N} \delta_i. \tag{2b}$$

We now wish to minimize the quantity Q with which we denote the mean-square distance between members of F:

Minimize Q

$$= \frac{1}{M(M-1)} \sum_{m=1}^{M} \sum_{p=1}^{M} \sum_{i=1}^{N} \left[\sum_{j=1}^{N} a_{ij}(f_{mj} - f_{pj}) \right]^2. \tag{3}$$

Before proceeding with the minimization, the quadratic form Q may be simplified. Expanding the squared expression as a double sum and interchanging the order of summations result in (4), where the bracketed quantity is a constant that depends only on the known samples of class F;

$$Q = \frac{M}{(M-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{s=1}^{N} a_{ij} a_{is} \cdot \left[\frac{1}{M^2} \sum_{m=1}^{M} \sum_{p=1}^{M} (f_{mj} - f_{pj})(f_{ms} - f_{ps}) \right]. \tag{4}$$

The above equation may be simplified by recognizing that the constant (the bracketed expression) is an element of the sample covariance matrix U.

$$2u_{js} = 2u_{sj} = \frac{1}{M^2} \sum_{m=1}^{M} \sum_{p=1}^{M} (f_{mj} - f_{pj})(f_{ms} - f_{ps}), \tag{5a}$$

$$= 2(\overline{f_j f_s} - \overline{f_j}\,\overline{f_s}). \tag{5b}$$

Hence the quantity Q may be written as in (6). Q may also be expressed conveniently in matrix notation (7), where the vector $a_i = (a_{i1}, a_{i2}, \cdots, a_{iN})$, and the prime denotes the transpose vector:

$$Q = \frac{2M}{(M-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} \sum_{s=1}^{N} a_{ij} a_{is} u_{js}; \tag{6}$$

$$Q = \frac{2M}{(M-1)} \sum_{i=1}^{N} a_i U a_i'. \tag{7}$$

The constraint (2b) can also be expressed in matrix notation, and it is given in (8), where I is the identity matrix;

$$1 = \prod_{i=1}^{N} a_i I a_i' = \prod_{i=1}^{N} \delta_i. \tag{8}$$

Minimization of Q, subject to the constraint given in (8), may now be carried out by the method of Lagrange multipliers.[2] The quantity Q and the constraint are differentiated with respect to a_i and their linear combination is equated to zero in (9), where the constant $2M/(M-1)$ is lumped into the Lagrange multiplier λ.

$$\sum_{i=1}^{N} a_i [U - \lambda(\prod_{l \neq i} a_l I a_l') I]\, da_i = 0. \tag{9}$$

Since da_i is arbitrary, every term in the above sum must be independently zero. Since the expression in parentheses is a constant that depends on i, the simplifications indicated in (10b) and (10c) can be made.

$$a_i [U - \lambda(\prod_{l \neq i} a_l I a_l') I] = 0; \qquad i = 1, 2, \cdots, N. \tag{10a}$$

$$a_i [U - \lambda_i I] = 0; \qquad i = 1, 2, \cdots, N, \tag{10b}$$

where

$$\lambda_i = \lambda(\prod_{l \neq i} a_l I a_l') = \lambda \frac{1}{\delta_i} \tag{10c}$$

from (8). For every eigenvalue λ_i for which (11) has a solution, there is a corresponding eigenvector a_i which

[1] G. Birkhoff and S. MacLane, "A Survey of Modern Algebra," The Macmillan Co., New York, N. Y.; 1953.

[2] F. B. Hildebrand, "Methods of Applied Mathematics," Prentice-Hall, Inc., Englewood Cliffs, N. J.; 1952.

column of the desired linear transformation A of observation space.

$$| U - \lambda_i I | = 0. \tag{11}$$

nce the covariance matrix is positive definite, its nvalues are real, and the corresponding eigenvectors orthogonal.[2] However, we have still not completely ed for the transformation coefficients, for the magni-s of the eigenvectors $\sqrt{\delta_i}$ are not known yet.

order to determine the δ_i's, we multiply (10b) by and sum over all i to obtain (12). This quantity is mean-square distance Q which we wish to minimize, ect to the constraint given in (8). The quantity to minimized is L, given in (13). Through use of the hod of Lagrange multipliers we obtain (14), where γ arbitrary constant.

$$Q \propto \sum_{i=1}^{N} a_i U a_i' = \sum_{i=1}^{N} \lambda a_i I a_i' = \sum_{i=1}^{N} \lambda_i \delta_i; \tag{12}$$

$$L = \sum_{i=1}^{N} \lambda_i \delta_i - \gamma\left(\prod_{i=1}^{N} \delta_i - 1\right); \tag{13}$$

$$- \gamma \prod_{l \neq i} \delta_l = 0 = \lambda_i - \gamma \frac{1}{\delta_i}; \; i = 1, 2, \cdots, N. \tag{14}$$

mposing the constraint of (15), we can solve for the stant γ, given in (16).

$$\prod_{i=1}^{N} \delta_i = 1 = \frac{\gamma^N}{\prod_{i=1}^{N} \lambda_i}; \tag{15}$$

$$\gamma = \left(\prod_{i=1}^{N} \lambda_i\right)^{1/N}. \tag{16}$$

stituting γ into (14), we can solve for δ_i as follows:

$$\delta_i = \frac{1}{\lambda_i}\left(\prod_{i=1}^{N} \lambda_i\right)^{1/N}. \tag{17}$$

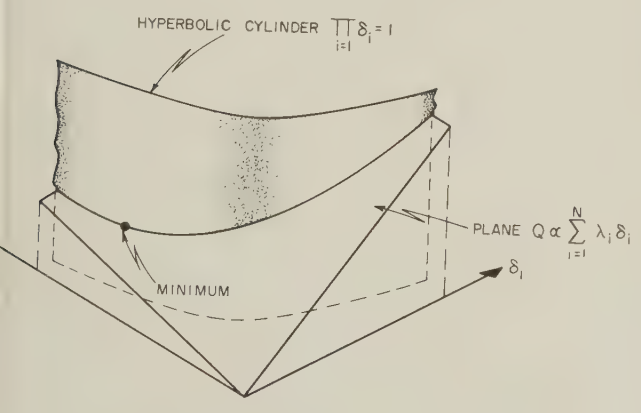


Fig. 2—Geometric interpretation of minimization.

he simple geometrical argument illustrated in Fig. 2 fies that the solution thus obtained indeed minimizes s a function of the δ_i's, Q is a plane [see (12)] whose rsection with the hyperbolic cylinder (the constraint) only one point of zero derivative. This point is a imum.

Thus the optimum of the class of metrics we considered, which minimizes the mean-square distance between members of the class F, is obtained by measuring the Euclidean distance on a linear transformation of the observation space. The transformation consists of a rotation and a diagonal transformation. Columns of the rotation transformation matrix R are unit eigenvectors of the sample covariance matrix U, and elements of the diagonal transformation D are inversely proportional to the square roots of the corresponding eigenvalues. The function S, defined by (1), is now given by (18) and denoted by $f(x)$, to simplify the notation. The bar denotes averaging over the given sample points f_m.

$$S(x, \{f_m\}) = \overline{(x - f_m)R \; DI \; D'R'(x - f_m)'} = f(x). \tag{18}$$

Similar functions may be developed for each of the classes. Suppose that there are only two classes F and G, and that the mean-square distances between x and members of the respective classes, as measured by two different metrics, are denoted by $f(x)$ and $g(x)$. The decision whether x is a member of F or G is made by comparison of $f(x) - g(x)$ with a threshold T, as expressed in (19). The locus of points for which $f(x) - g(x) = T$, a constant, serves as the boundary between the two regions into which the observation space is partitioned.

$$f(x) - g(x) \gtrless T. \tag{19}$$

Relationship to Decision Theory

The relationship between decisions based on likelihood ratios and those made by the method described thus far are discussed in this section. It will be shown that if the classes are Gaussian processes with unknown but, in general, different means and variances, then S defines contours of equal *a posteriori* probabilities. That is, $S(x, \{f_m\})$ measures the mean-square distance by a non-Euclidean metric between a point x and M members of an ensemble of points $\{f_m\}$, and is a measure of the probability that x belongs to class F.

Fixed values of S correspond to contours of equal *a posteriori* probability. The ratio of *a posteriori* probabilities is proportional to the likelihood ratio, the logarithm of which will be shown to be equal to $f(x) - g(x)$ above.

Consider the situation where an arbitrary event x may be a member of only one of two classes F or G. The likelihood ratio that x belongs to F rather than to G is expressed by the ratio of *a posteriori* probabilities in (20), which may be simplified by Bayes Rule:

$$\frac{p(F/x)}{p(G/x)} = \frac{p(x/F)p(F)/p(x)}{p(x/G)p(G)/p(x)} \propto \frac{p_F(x)}{p_G(x)} = l(x). \tag{20}$$

The likelihood ratio is thus proportional to the ratio of the two joint probability densities of the two Gaussian processes. In the event that membership in either of the two classes is *a priori* equally likely, the proportionality becomes an equality.

Now let us examine the probability density $p_F(x)$, a factor of the likelihood ratio $l(x)$. For the multivariate Gaussian process, the joint probability density is given by (21), where U is the covariance matrix of F, and $[U_{rs}]$ is the cofactor of the element with like subscripts in the covariance matrix.[3] It should be noted that $|U_{rs}|/|U|$ is an element of U^{-1}.

$$p_F(x_1, x_2, \cdots, x_N) = \frac{1}{(2\pi)^{N/2} \mid U \mid^{1/2}} \cdot \exp\left[-\tfrac{1}{2} \sum_{r=1}^{N} \sum_{s=1}^{N} \frac{\mid U_{rs} \mid}{\mid U \mid} (x_r - m_r)(x_s - m_s) \right], \quad (21a)$$

$$= \frac{1}{(2\pi)^{N/2} \mid U \mid^{1/2}} \cdot \exp\left[-\tfrac{1}{2} \sum_{r=1}^{N} \sum_{s=1}^{N} \mid U_{rs}^{-1} \mid (x_r - m_r)(x_s - m_s) \right]. \quad (21b)$$

Contours of constant joint probability density occur for those values of x for which the argument of the exponential is constant. The exponent expressed in matrix notation is

$$\text{exponent} = [-\tfrac{1}{2}(x - m_x)U^{-1}(x - m_x)']. \quad (22)$$

It will be recalled that one of the operations on the set of points $\{f_m\}$ which the optimum metric performed was a rotation, expressible by an orthogonal matrix R. This is a pure rotation (an orthonormal transformation), where columns of R are unit eigenvectors of the covariance matrix U.

Let y be a new variable obtained from x by (23). Substituting (23b) into (22), (24) is obtained as follows:

$$y = xR, \quad (23a)$$

$$x = yR^{-1}; \quad (23b)$$

$$\text{exponent} = [-\tfrac{1}{2}(y - m_y)R^{-1}U^{-1}[R^{-1}]'(y - m_y)']. \quad (24)$$

Since R is orthogonal, the special property of orthogonal matrices that $R^{-1} = R'$ may be utilized to simplify (24). This yields

$$\text{exponent} = [-\tfrac{1}{2}(y - m_y)R'U^{-1}R(y - m_y)']. \quad (25)$$

Furthermore, since columns of R are eigenvectors of the covariance matrix U, the matrix R must satisfy (26a), where Λ is the diagonal matrix of eigenvalues of $\mid U - \lambda_n I \mid = 0$.

$$R'[U - \Lambda]R = 0; \qquad R'UR = R'\Lambda R = \Lambda. \quad (26a)$$

$$\Lambda = \begin{bmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_N \end{bmatrix}. \quad (26b)$$

By taking the inverse of both sides of (26a) and employing again the special property of orthogonal matrices, (27) may be obtained. This latter expression, when substituted into (25), produces the result stated in (28).

$$R'U^{-1}R = \Lambda^{-1}; \quad (27)$$

$$\text{exponent} = [-\tfrac{1}{2}(y - m_y)\Lambda^{-1}(y - m_y)']. \quad (28)$$

The quadratic form of (28) expresses the fact that contours of constant probability density are ellipsoids with centers at m_y; the direction of the principal axes are along eigenvectors of the covariance matrix, and the diameters are equal to the corresponding eigenvalues. This result can be shown in a more familiar form by converting the quadratic form of (28) to a sum, as shown in (29), in which y_n is the coordinate of y in the direction of the nth eigenvector, and m_n is the mean of the ensemble in the same direction:

$$\text{exponent} = \left[-\tfrac{1}{2} \sum_{n=1}^{N} \frac{(y_n - m_n)^2}{\lambda_n} \right]. \quad (29)$$

An expression of identical appearance can be derived from the exponent of the joint probability density of class G. The differences between the two exponents exist in 1) the directions of their eigenvectors, 2) the numerical magnitude of their eigenvalues, and their ensemble means. Denoting the exponents in the two probability densities by $e_f(x)$ and $e_g(x)$, the logarithm of the likelihood ratio may be written as in (30), where K is a constant which involves the ratio of *a priori* probabilities and the ratio of determinants of the two covariance matrices:

$$\log l(x) = K + e_f(x) - e_g(x). \quad (30)$$

Now we will show that $e_g(x)$ is proportional to $g(x)$ and $e_f(x)$ is proportional to the previously derived $f(x)$ and thus prove that, for the special case when classes are Gaussian processes, partitioning the observation space into regions by (19) is identical to that achieved through decision theory. In accordance with the foregoing remarks we wish to prove

$$f(x) = \overline{(x - f_m)R\ D\ I\ D'R'(x - f_m)'} \propto e_f(x) \propto \sum_{n=1}^{N} \frac{(y_n - m_n)^2}{\lambda_n}. \quad (31)$$

Recognizing (from the definition of D) that $D\ I\ D' = \Lambda^{-1}$ and making the change of variables, $y = xR$, (32) is obtained, where F_m is the transformed vector f_m.

$$f(x) = \overline{(y - F_m)\Lambda^{-1}(y - F_m)'} \propto \sum_{n=1}^{N} \frac{(y_n - m_n)^2}{\lambda_n} \propto e_f(x). \quad (32)$$

Writing this as a sum and interchanging the order of averaging and summation yields

$$\overline{\sum_{n=1}^{N} \frac{(y_n - F_{mn})^2}{\lambda_n}} = \sum_{n=1}^{N} \frac{\overline{(y_n - F_{mn})^2}}{\lambda_n} \propto \sum_{n=1}^{N} \frac{(y_n - m_n)^2}{\lambda_n}. \quad (33)$$

By expanding the square, and adding and subtracting $(\overline{F_n}$

[3] W. B. Davenport and W. L. Root, "Random Signals and Noise," McGraw-Hill Book Co., Inc., New York, N. Y.; 1958.

m each term of the numerator, we obtain (34), where $= m_n$ and $\overline{F_n^2} - (\overline{F_n})^2 = \sigma_n^2 = \lambda_n$. Thus the proportion-y of $f(x)$ and $e_f(x)$ is established. It is now readily seen t $f(x) - g(x) = e_f(x) - e_g(x)$, and contours of constant elihood ratio $l(x)$ are identical to contours of constant $) - g(x)$.

$$\overline{\frac{(y_n - F_{mn})^2}{\lambda_n}} = \sum_{n=1}^{N} \frac{(y_n - m_n)^2 + \lambda_n}{\lambda_n}$$

$$\propto \sum_{n=1}^{N} \frac{(y_n - m_n)^2}{\lambda_n} \qquad (34)$$

Experimental Results

An experimental program was devised to verify the hnique discussed above. This program, consisting of chine recognition of spoken numerals, was of suffi-ntly high dimensionality to exhibit not only the success the method, but also the ease with which it can be plemented on present-day computers. The numerals ero" through "nine" were spoken a number of times ten male speakers drawn from the Northeast section the United States. Each utterance of a numeral formed basis of a sample point in a high-dimensional space. e utterances were represented as vectors by means an eighteen-channel Vocoder.[4] The Vocoder is a set eighteen stagger-tuned band-pass filters that produce the filter outputs the "instantaneous" frequency spec-m of the utterance, as a function of time.

An example of the output of the Vocoder is given in g. 3, in which the ordinate is frequency and the abscissa time. The intensity of the spectrum at a given time and quency is indicated by the grayness of the sonograph the prescribed time and frequency point. In order to m a vector from the sonograph, the spectrum is sampled 20-msec intervals in each of the frequency channels. e array of sample heights is a vector, an example of ich is given beneath the sonograph of Fig. 3.

Because the duration of utterances of numerals varies th the numeral and with the speaker, all records were rmalized in time by multiplication by scale factors, d the sonographs were resampled to produce twenty e intervals per *utterance*. The $18 \times 20 = 360$-dimen-nal vector was then augmented by the scale factor as additional dimension. This relatively unsophisticated proach to normalization was adopted because of the ited scope of the experiment.

Recognition of membership in classes involves learning a small number of samples, at first, and then increasing sample size while testing unlabeled points at each ge of the experiment to see whether they are correctly ssified. Learning to recognize spoken numerals, in this periment, consisted of forming the covariance matrices each of the ten sets of given sample words, and of solving the optimum transformations which maximally clus-ed each set. A new word was then classified as a member

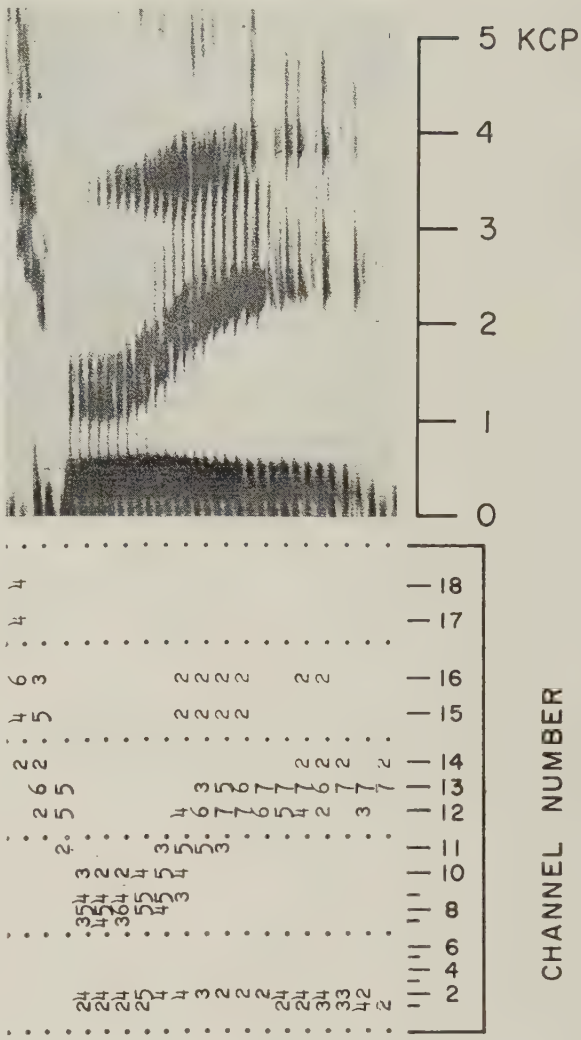
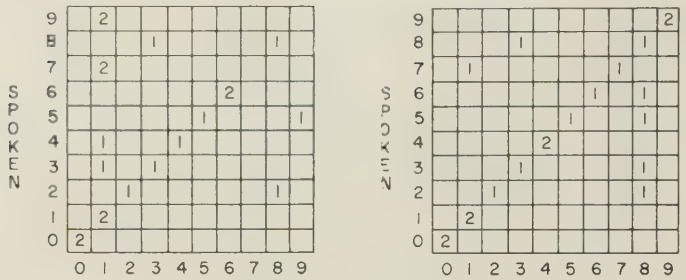


Fig. 3—Vocoder representation of the spoken word "three."

of one of the spoken numerals by computing the ten different mean-square distances—the $f(x)$'s of (31)—and then deciding that the new word belonged to that class of numerals to which the corresponding $f(x)$ was smallest.

A typical result which demonstrates improvement in the machine's performance as the number of known, labeled examples of spoken digits is increased, is illustrated in Fig. 4. This figure contains four confusion matrices constructed for the cases where numeral recognition was learned from 3, 4, 7, and 9 examples of each of the ten classes of digits. The ordinate of a cell in the matrix signifies the digit which is spoken, the abscissa denotes the decision of the machine, and the number in the cell indicates the number of instances in which the stated decision was made. The number 1 in row 6 and column 8 of Fig. 4(c), for example, denotes the fact that in one instance a spoken digit 6 was recognized as an 8. Note that the error rate decreases as the number of known examples of classes is increased. For the 9 examples per class, no errors were made. This result is particularly interesting in view of the fact that the digits which were tested were spoken by persons not included among those whose words were used as examples.

[4] The Vocoder data used were made available through the rtesy of the AF Cambridge Res. Ctr.



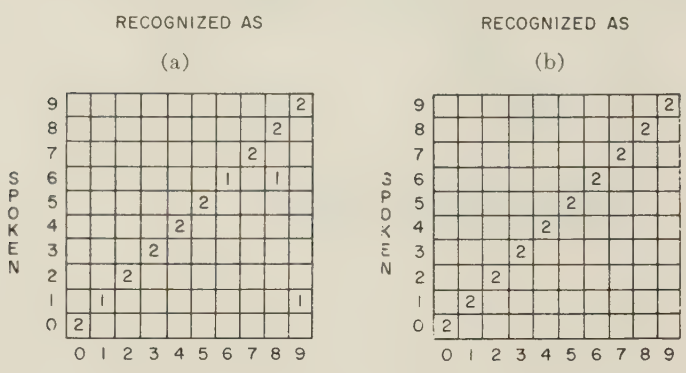


Fig. 4—Confusion matrices illustrating the process of learning spoken numeral recognition. (a) 3 examples per class, error rate 45 per cent. (b) 4 examples per class, error rate 30 per cent. (c) 7 examples per class, error rate 10 per cent. (d) 9 examples per class, error rate 0 per cent.

Concluding Remarks

A geometrical approach to recognition of membership in classes is presented. The approach consists of postulating a new decision rule which involves the comparison of mean-square distances between the input vector and members of the different classes. Distance to each class in this comparison, is measured by different metrics generated by minimization problems which cluster sample points of each class separately. The metrics thus obtained are shown to equate the new decision rule with Bayes Rule when classes are assumed Gaussian processes, and when the sample covariance matrix approximates the covariance matrix of the process. This is, in a sense, what is implied by the stated assumption that the examples of the classes are representative.

Several extensions of the method described were obtained by consideration of larger classes of metrics involving highly nonlinear operations. In addition, other methods of generating metrics to be used in the decision rule were considered. In particular, a number of solutions were obtained for generating metrics which minimize distances between members of the same class, while separating members of different classes. A discussion of metrics of this type and their decision-theoretical implications will be the subject of another paper.

Correction

Thomas Kailath, author of "Correlation Detection of Signals Perturbed by a Random Channel," which appeared on pages 361–366 of the June, 1960 issue of these Transactions, has brought the following changes to the attention of the Editor.

On page 365, column 1, the first unnumbered equation after (27) should read

$$\mathbf{x}_t^{(k)} = \left| \mathbf{x}_t \vdots \hat{\mathbf{x}}_t \right| = \left| x_0^{(k)} x_1^{(k)} \cdots \vdots \hat{x}_0^{(k)} \hat{x}_1^{(k)} \cdots \right| .$$

The second unnumbered equation after (27) should read

$$\begin{bmatrix} \mathbf{z} \\ \cdots \\ \hat{\mathbf{z}} \end{bmatrix} = \left[\begin{array}{c:c} \mathbf{A} & -\hat{\mathbf{A}} \\ \cdots & \cdots \\ \hat{\mathbf{A}} & \mathbf{A} \end{array} \right] \begin{bmatrix} \mathbf{x} \\ \cdots \\ \hat{\mathbf{x}} \end{bmatrix} .$$

orrespondence

Close-Packed Double Error-recting Codes on P Symbols*

et p be a positive integer. Let E_p^d be space of all d-tuples whose entries are en from the set 0, 1, 2, $\cdots$, $p-1$. define the distance between two points E_p^d as the number of places in which disagree. A close-packed double error-ecting code is a subset S of E_p^d such any two points of S are no closer than each other and such that any point of is within 2 of some (hence, a unique) nt of S.

or $p = 2$, Shapiro and Slotnik[1] have wn such a code exists only for the trivial = 5. For $p = 3$, such a code for $d = 11$ found by Golay,[2] and Lee[3] has shown to be the only possible d. Lee[3] has also wn there are no such codes for $p = 4$. We shall treat the case $p = 5$ with some ner general methods that we believe extend to other p. Now if such a code sts, it is easy to see that the space E_5^d aks up into spheres of radius two about points of S. That is, all such spheres disjoint and together they fill E_5^d. e number of points in such a sphere is $- 4d + 4^2d(d-1)/2$ and the number of h spheres is

$$\frac{5^d}{1 + 4d + \frac{4^2 d(d-1)}{2}}. \tag{1}$$

w, (1) is an integer, so

$$1 + 4d + \frac{4^2 d(d-1)}{2} = 5^k; k \leq d.$$

we let

$$Z = 4d - 1 \tag{2}$$

n the above equation becomes

$$Z^2 + 1 = 2 \cdot 5^k. \tag{3}$$

Now, (3) is clearly satisfied for $k = 0$ nd 2. But tracing through (2) and (3) see that these correspond to $d = 0$, 1, 1 2, respectively, which are impossible double error-correcting codes. We wish nvestigate $k > 2$, and our result is that re are no further solutions. In (3), Z is iously odd, $Z = 2n + 1$, say. Eq. (3) omes

$$n^2 + (n+1)^2 = 5^k. \tag{4}$$

We shall, from this point on, restrict k and n to solutions of (4). We show first Lemma 1.

Lemma 1:

$$n \equiv 28 \quad \text{or} \quad -29 \pmod{125}. \tag{5}$$

Proof: If n runs through the integers 0, 1, 2, 3, and 4, then $n^2 + (n+1)^2$ runs through 1, 0, 3, 0, 1 (mod 5). Thus we see $n \equiv 1$ or 3 (mod 5). If $n \equiv 1 \pmod 5$, $n = 5j + 1$. Substituting this in (4), we obtain

$$5^{k-1} = 10j^2 + 6j + 1,$$

or $j \equiv -1 \pmod 5$, $j = 5l - 1$, and hence $n = 25l - 4$. Substituting this in (4), we obtain $5^{k-2} = 50l^2 - 14l + 1$ or $l \equiv -1 \pmod 5$, $l = 5t - 1$. So $n = 125t - 29$, *i.e.*, $n \equiv -29 \pmod{125}$.

The corresponding computation shows that if we start with $n \equiv 1 \pmod 5$, we obtain $n \equiv 28 \pmod{125}$.

These results constitute a sieve on n; we really want a sieve on k. But we shall see that Lemma 1 is useful in obtaining conditions on k.

We now pass to the Gaussian integers about which we need the following facts which we shall quote without proof. (See, for example, Landau[4] or Zariski and Samuel.[5]) The integers are the numbers of the form $a + bi$ where a and b are natural integers. Unique factorization holds. $2 + i$ and $2 - i$ are primes. The only units are $+1$, -1, $+i$, and $-i$. Let us factor (4):

$$[(n+1) + ni][(n+1) - ni] = (2+i)^k(2-i)^k, \tag{6}$$

with the possible inclusion of some units. We now wish to find the factors of $(n+1) + ni$. Should $(n+1) + ni$ have as factors both $2 + i$ and $2 - i$, it would have a factor of 5 and hence its real and imaginary parts, $(n+1)$ and n, would both be divisible by 5, which is impossible. This shows there are eight possibilities:

$$(n+1) + ni = \mu^k, -\mu^k, i\mu^k, -i\mu^k, \mu^{-k}, -\mu^{-k}, i\mu^{-k}, \text{or } -i\mu^{-k}, \tag{7}$$

where $\mu = 2 + i$. Now, let

$$\beta_j = \frac{\mu^j + \mu^{-j}}{2}. \tag{8}$$

Summing the geometrical series shows the generating function of the β_j to satisfy

$$(1 - 4Z + 5Z^2) \sum_0^\infty \beta_j Z^j = 1 - 2Z. \tag{9}$$

This yields a recursion formula for β_j,

$$\beta_0 = 1, \quad \beta_1 = +2,$$

$$\beta_j = 4\beta_{j-1} - 5\beta_{j-2}; \quad j \geq 2. \tag{10}$$

On the other hand, if n and k constitute a solution of (11) we can compute β_k from (7) and (8). This yields the four possibilities,

$$\beta_k = \pm n, \qquad \pm(n+1). \tag{11}$$

From (11) and Lemma 1, we see that a solution of (4) entails

$$\beta_k \equiv \pm 28 \text{ or } \pm 29 \pmod{125}. \tag{12}$$

Now we must compute β_j from (10), watching for the occurrence of (12). It clearly suffices to compute (10) (mod 125). If we compute the β_j for any integral modulus, the β_j must be cyclic since the residue class is finite and the repetition of any consecutive pair of β_j's means the recursion repeats. Table I represents the computation of the β_j (mod 125).

TABLE I
β_j (mod 125)

j	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
β_j	1	2	3	2	−7	−38	8	−28	−27	32	13	17	3	52	−57	12	−42

j	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32
β_j	22	48	−43	−37	−58	−47	−23	18	62	33	−53	−2	7	38	−8	28

Applying the criteria of (12) to Table I, we see that

$$k \equiv 7 \pmod{25}. \tag{13}$$

We shall now give a demonstration due to H. F. Mattson that excludes the possibility of (13) and hence the existence of close-packed double error-correcting codes on five symbols. The basic idea is that if $2 \cdot 5^k - 1$ is a square, then it is a quadratic

* Received by the PGIT, May 11, 1960; revised uscript received, August 30, 1960.

[1] H. S. Shapiro and D. L. Slotnik, "On the mathe-ical theory of error-correcting codes," *IBM J. Dev.* vol. 3, pp. 25–34; January, 1959.

[2] M. J. E. Golay, "Notes on digital coding," c. IRE, vol. 37, p. 657; June, 1949.

[3] C. Y. Lee, "Some properties of nonbinary r-correcting codes," IRE TRANS. ON INFORMA-N THEORY, vol. IT-4, pp. 77–82; June. 1958.

[4] E. Landau, "Elementary Number Theory," Chelsea Publishing Co., New York, N. Y. (reprint); 1958.

[5] O. Zariski and P. Samuel, "Commutative Algebra," D. Van Nostrand Co., Inc., New York, N. Y.; 1958.

residue for any modulus, an immediate consequence of the definition of quadratic residue. Let q be a prime $\neq 5$ and t be the order of 5 (mod q), *i.e.*, some power of 5 is congruent to 1, *e.g.*, 5^{q-1} by Fermat's theorem, and we let t be the least positive integer such that $5^t \equiv 1 \pmod{q}$. Then

$$5^{jt+r} - 1 \equiv 5^r - 1 \pmod{q}$$

$$j = 0, 1, 2, \cdots .$$

Thus the numbers of the form $5^{jt+r} - 1$, $j = 0, 1, 2, \cdots$ are either all quadratic residues or all quadratic nonresidues. To show the impossibility of (13), it suffices, therefore, to find a prime q such that $5^7 - 1$ is a nonresidue and 5 is of order 25 (mod q). We shall show $q = 101$ suffices. Some powers of 5 (mod 101) are

$$5^5 \equiv -6, \quad 5^7 \equiv -49, \text{ and } 5^{25} \equiv 1.$$

This shows that 5 is of order 25 and that $2 \cdot 5^7 - 1 \equiv 2 \pmod{125}$. But 2 is a nonresidue of primes of the form $8n + 5$ (see Zarinski and Samuel[5]) and, in particular, of 101. This completes our proof that there exist no close-packed double error-correcting codes on 5 symbols.

Carl Engelman
The Mitre Corp.
Bedford, Mass.

Matched Filters for Multiple Processes*

By a multiple process is meant a q-dimensional (real-valued) vector stochastic process. As is well-known, a matched filter is defined for one-dimensional processes as one that minimizes the noise-to-signal ratio.[1] Now, in the multiple process case we can define noise-to-signal ratio in several ways and it is the purpose of this note to examine the corresponding matched filters.

A filter (linear filter) now corresponds to a m-by-q matrix function $W(t)$. We have q inputs and m outputs related by

$$Y(t) = \int_{-\infty}^{\infty} W(t - \sigma)\, X(\sigma)\, d\sigma$$

where $X(\sigma)$ is the q-dimensional input process and $Y(t)$ the m-dimensional output. For physical realizability it is necessary that

$$W(t) = 0 \text{ [zero matrix] for } t < 0,$$

and we shall assume this in what follows. Suppose now $X(t)$ is composed of signal and noise

$$X(t) = S(t) + N(t)$$

where $S(t)$ is the q-dimensional determinate signal, and $N(t)$ is the q-dimensional noise (stochastic) process. It is convenient to think of $S(t)$ and $N(t)$ as q-by-1 matrices. The covariance matrix function $R(s, t)$ of the noise process is defined by

$$R(s, t) = E[N(s)\, N(t)^*],$$

the asterisk denoting the adjoint. The filter $W(t)$ acting on $X(t)$ being linear, we have both noise and signal "terms" in $Y(t)$. The response at an instant of time T is given by

$$Y(t) = \int_0^T W(T - t)\, S(t)\, dt + \int_0^T W(T - t)\, N(t)\, dt.$$

The noise-to-signal ratio N/S in $Y(t)$ can be defined in various ways. First, if we consider them as vectors, the squared magnitude of a vector being the sum of the squares of the components, we can define

$$(N/S)_1 = \frac{\text{average squared magnitude of noise}}{\text{squared magnitude of signal at time } T}.$$

Again, we can also define it as

$$(N/S)_2 = \frac{\text{average squared magnitude of noise}}{\text{square of the sum of the signal components at time } T}.$$

Also as

$$(N/S)_3 = \text{sum of noise-to-signal ratios of each component.}$$

We shall show that all of these notions lead to the same matched filter. For this, it is convenient to introduce the notion of a linear operator R on the (Hilbert) space $\mathcal{H}$ of q-dimensional square-integrable functions defined on $[o, T]$. Thus, for any element g in $\mathcal{H}$, we define

$$Rg = h$$

where

$$h(s) = \int_0^T R(s, t)\, g(t)\, dt.$$

The inner product in $\mathcal{H}$ of any two elements h and g with components h_i, g_i will be denoted

$$[g, h] = \sum_{i=1}^{q} \int_0^T h_i(t)\, g_i(t)\, dt.$$

Let

$$h_i(t) = W_i(T - t), \quad 0 \leq t \leq T,$$

be the rows of $W(T - t)$. Then each h_i belongs to $\mathcal{H}$ and the average of the sum of the squares of the noise components of $Y(t)$ is easily seen to be

$$\sum_{i=1}^{m} [Rh_i, h_i].$$

For any choice h_i, let

$$\gamma_i = [h_i, S].$$

Then

$$(N/S)_1 = \frac{\sum_1^m [Rh_i, h_i]}{\sum_1^m \gamma_i^2}.$$

Keeping $\{\gamma_i\}$ fixed, suppose we vary $\{h_i\}$ in order to minimize the ratio $(N/S)_1$. Then it follows, as shown in a previous article,[2] that the minimum is given by

$$\frac{1}{[h, S]}$$

where h is the solution of the (integral) equation

$$Rh = S.$$

The optimal $\{h_i\}$ corresponding to this minimum is given by (within multiplicative scalar constants)

$$h_i = h, \quad i = 1, \cdots m.$$

In a similar manner, it can be shown that the same solution minimizes the other (N/S) ratios. The minima are given by

$$\min (N/S)_2 = \frac{1}{m} \frac{1}{[h, S]},$$

$$\min (N/S)_3 = m \frac{1}{[h, S]}.$$

The details of the solution when there is no element h in $\mathcal{H}$ such that

$$Rh = S$$

may be found in a previous paper.[2]

A. V. Balakrishnan
Space Tech. Labs., Inc.
Los Angeles, Calif.

* Received by the PGIT, July 22, 1960.

[1] G. L. Turin, "An introduction to matched filters," IRE Trans. on Information Theory, vol. IT-7, pp. 311–329; June, 1960.

[2] A. V. Balakrishnan, "Estimation and detection theory for multiple stochastic processes," *J. Math. Analysis and Applications*, vol. 1, pp. 1–12; January, 1961.

ıential Generation and Decoding ıe *P*-Nary Hamming Code*

veral recent papers[1–3] have shown how r sequential circuits operating over ɔdular field may be used to generate decode single- and multiple-error cting codes. This note is concerned the possibility of generating and dling *Hamming* codes[4] with the same of circuit, which has the feature of ring only as many delays as there checking digits, and also leads to a le decoder. Other papers[5–9] have been rned with sequential generation of -correcting codes, but have resulted rcuits with a larger number of delay nts than necessary, and have con- ed only the binary case.

ıe Hamming codes have the property the information digits pass through oder unchanged, and that the checking s are linear combinations of the mation digits. The conclusions of this stigation are: 1) The Hamming codes

$$\begin{bmatrix} y_1 \\ y_2 \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ \cdot \\ y_{m+k} \end{bmatrix} = \begin{bmatrix} h_{11} & 0 & & \cdots & 0 \\ h_{21} & h_{12} & 0 & \cdots & 0 \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ \cdot & & & & \\ h_{m+k,1} & \cdots & & & h_{m+k,m+k} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \cdot \\ \cdot \\ \cdot \\ x_m \\ 0 \\ \cdot \\ \cdot \\ 0 \end{bmatrix}$$

be generated by linear modular ential circuits whose order (*i.e.*, the ber of delay elements) equals the ber of check digits, but this circuit t be time-varying. 2) The decoder, h is the inverse of the coder, has a structure which is simply related to the coder. In the binary case, in fact, the coder is self-inverse, *i.e.*, the decoder and coder are identical. 3) A simple error-correction scheme may be used.

For the coder to be realizable as a sequential circuit, the check digits occur after the information digits which they check. For simplicity we have chosen to have the coded word consist of the information digits in the first m positions and the check digits in the last k positions. Hence the input to the coder will be the sequence $x_1x_2 \cdots x_m\ 0\ 0 \cdots 0$, where the number of zeros equals the number of check digits to be generated in the coder. The output of the coder will be

$$y_1y_2 \cdots y_my_{m+1}y_{m+2} \cdots y_{m+k},$$

where $y_i = x_i$ for $i \leq m$, and y_i equals the Hamming check digit for $i > m$.

Now, in a linear sequential circuit, the input-output relation may be represented in the following matrix form:

where

$$h_{ij} = h_{i-j} \qquad (1)$$

if and only if the circuit is time-invariant. Note that the transmission matrix[10] $[h_{ij}]$ is lower triangular, a necessary condition for it to represent a realizable network. Since the coding scheme requires that $y_1 = x_1, y_2 = x_2, \cdots, y_m = x_m$, it is apparent that

$$h_{ij} = \delta_{ij} \qquad 1 \leq i \leq m.$$

Also, since $x_j = 0$ for $j = m + 1, m + 2, \cdots, m + k$, we can set $h(i, j) = \phi_{ij}$ for $i, j = m + 1, m + 2, \cdots, m + k$, where ϕ_{ij} is arbitrary. However, physical realizability requires that $\phi_{ij} = 0$ for $j > i$ and the existence of an inverse for H necessitates that $\phi_{ii} \neq 0$. Hence we find that the transmission matrix H of the coder may be partitioned in the following manner:

$$\mathbf{H} = \begin{bmatrix} \mathbf{I}_m & \mathbf{O} \\ \mathbf{C} & \mathbf{\Phi}_k \end{bmatrix}$$

where $\mathbf{I}_m$ is an $m \times m$ unit matrix, $\mathbf{\Phi}_k$ is an arbitrary $k \times k$ nonsingular triangular matrix, and $\mathbf{O}$ is the null matrix. The matrix $\mathbf{C}$ is the check matrix whose elements are determined by the nature of the checking scheme to be employed.

To demonstrate that H cannot be the transmission matrix of a time-invariant circuit we consider first an example for $p = 2$, $m = 4$, $k = 3$; the corresponding transmission matrix is

$$\mathbf{H} = \left[\begin{array}{cccc:c} 1 & 0 & 0 & 0 & \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & \\ 0 & 0 & 0 & 1 & \\ \hdashline h_{40} & h_{41} & h_{42} & h_{43} & \\ h_{50} & h_{51} & h_{52} & h_{53} & \Phi \\ h_{60} & h_{61} & h_{62} & h_{63} & \end{array}\right].$$

Since $h_{10} = 0$, then $h_{43} = h_{54} = h_{65} = 0$ to satisfy (1). Likewise $h_{30} = 0$ and $h_{20} = 0$ forces the condition $h_{41} = h_{42} = h_{52} = h_{53} = h_{63} = 0$. Hence the 4th column of the check matrix $\mathbf{C}$ is zero, which implies that the check digits are independent of the information digit $x(3)$, implying that this digit is not checked. In general, a fixed system would require $n - k - 1$ zeros, where $n = m + k$, following the nonzero elements on the major diagonal. In order that the nth digit be checked, there must be at least one nonzero element in the last k rows of the mth column. These are the k elements following the nonzero element on the major diagonal, and thus $n - k - 1 < k$. Using the relation that $(p - 1)n = p^k - 1$ for close-packed codes (including the Hamming code) it is found that

$$2k > \frac{p^k - 1}{p - 1} - 1$$

$$= p + p^2 + \cdots + p^{k-1}, \qquad (2)$$

which cannot be satisfied for any $p > 3$ or $k > 2$. Thus, a time-invariant coder leaves some digits unchecked, and hence, the Hamming coder cannot be time-invariant, except for two trivial cases of only one information digit each.

The simplest coder and decoder result when $\Phi = I_k$, a $k \times k$ unit matrix. Then we obtain the transmission matrix of the coder

$$\mathbf{H} = \begin{bmatrix} \mathbf{I}_m & \mathbf{O} \\ \mathbf{C} & \mathbf{I}_k \end{bmatrix}.$$

Now the decoder has a transmission matrix $\mathbf{H}_d$ which is the inverse of the transmission matrix of the coder. Hence, owing to the selection of $\Phi = I_k$ we find that

$$\mathbf{H}_d = \mathbf{H}^{-1} = \begin{bmatrix} \mathbf{I}_m & \mathbf{O} \\ -\mathbf{C} & \mathbf{I}_k \end{bmatrix}.$$

Received by the PGIT, June 20, 1960; revised ıscript received, August 15, 1960. This work was ɔrted by the Natl. Science Foundation under ract No. G-9780. Publication was assisted by Marcellus Hartley Fund.

D. A. Huffman, "A linear circuit viewpoint on -correcting codes," IRE Trans. on Information ɔry, vol. IT-2, pp. 20–28; September, 1956.
3. Elspas, "The theory of autonomous linear ntial networks," IRE Trans. on Circuit ɔry, vol. CT-6, pp. 45–60; March, 1959.
T. E. Stern and B. Friedland, "Application of lar sequential circuits to single error-correcting y codes," IRE Trans. on Information ɔry, vol. IT-5, pp. 114–123; September, 1959.
R. W. Hamming, "Error detecting and cor- g codes," *Bell Sys. Tech. J.*, vol. 29, pp. 147–160; , 1950.
N. M. Abramson, "A class of systematic codes ɔn-independent errors," IRE Trans. on Infor- ɔn Theory, vol. IT-5, pp. 150–157; December,
P. Fire, "A Class of Multiple-Error-Correcting y Codes for Non-Independent Errors," Sylvania ronic Systems Rept. presented at AIEE Fall ing, Chicago, Ill.; October, 1959.
C. M. Melas, "Ayclic code for double error ction," *IBM J. Res. and Dev.* vol. 4, pp. 366; July, 1960.
. E. Meggitt, "Error correcting codes for cting bursts of errors," *IBM J. Res. and Dev.*; ; pp. 329–334; July, 1960.
C. M. Melas, "A new group of codes for cor- n of dependent errors in data transmission," *J. Res. and Dev.*, vol. 4, pp. 58–65; January,

[10] B. Friedland, "A technique for the analysis of time-varying sampled data systems," *Trans. AIEE*, vol. 73, pt. 2, pp. 407–414; January, 1957.

It is seen that the decoder and coder are identical in structure, the only difference being that the matrix $-\mathbf{C}$ of the decoder is the additive inverse, modulo p, of the corresponding coder matrix. In the binary case the coder and decoder are identical, since $\mathbf{C} = -\mathbf{C}$, modulo 2.

The check matrix $\mathbf{C}$ of the coder may be obtained using the Hamming check rules[4], and is not unique. The only rule to be followed in constructing $\mathbf{C}$ is that $c_{ij} \neq 0$ if and only if the ith check digit checks the jth information digit. However, by adapting Huffman's error-correcting technique[1,3] we can obtain a very simple decoding system. The principle of this technique is the choice of a unit response whose $m+1, m+2, \cdots, m+k$ symbols uniquely identify the position in which the error has occurred. But, since in the Hamming code a single error does not propagate to the other information digits at the input to the decision mechanism, only one information digit must be corrected. Errors in the check digits alone need not be corrected.

Each single error in an information digit will result in a distinct sequence in the check digits (at the input to the decision mechanism). This sequence will contain at least two nonzero elements (since errors in the check digits will produce sequences containing a single nonzero element). An error of magnitude $+q$ ($q < p$) will produce a check sequence such that each digit is q times (mod p) the corresponding digit for an error of $+1$. An error correction sequence is required that contains consecutively all combinations of the check digits for single errors of $+1$ in the information digits. This sequence differs from the maximal length sequence of Stern and Friedland[3] in that it does not include combinations with only one nonzero element. It is possible to obtain such a sequence by considering only the first $(p^k - 1)/(p-1) - 1$ terms of a maximal length sequence, if the sequence is chosen such that the first term is unity and the next $k-2$ terms are zero. This choice is always possible by proper selection of the numerator polynomial of the generating pulse-transfer function.

Error correction is obtained by comparing the check digits entering the decision mechanism with the error sequence. The first comparison is made between the first k digits of the error sequence and the k check digits. The error sequence is then shifted one place for the second comparison, such that the second through $(k+1)$st digits of the error sequence are compared with the check digits. Subsequent comparisons are obtained by further shifts. If a match is obtained or if the check digits are multiples (digit by digit, mod p) of the correponding k digits of the error sequence on the rth comparison, then the digit in error is the $(n-k+1-r)$th. This digit is in error by $+q$, where q times the error sequence matches the check digits.

The $\mathbf{C}$ matrix must now be arranged in the order corresponding to the error sequence by an appropriate renumbering of the information digits. Thus, for example, the last column of the $\mathbf{C}$ matrix contains the first k digits of the error sequence, in order, and the first column contains the last k digits of the error sequence.

As an example, consider the case discussed above ($p = 2$, $n = 7$, $k = 3$). A possible sequence for this system would be $1\,0\,1\,1\,1\,0\,0, 10 \cdots$. The required sequence for the Hamming decoder is $1\,0\,1\,1\,1\,0$. Corresponding to this sequence, the information bits must be renumbered such that the second and fourth information bits of the standard Hamming word are interchanged. The $\mathbf{C}$ matrix becomes

$$\mathbf{C} = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 0 \\ 0 & 1 & 1 & 1 \end{bmatrix}.$$

If the input to the decision mechanism were $x_1x_2x_3x_4\,1\,0\,1$, it can be seen by comparing the check bits with the error sequence that x_4 is in error. If, on the other hand, the error sequence is shifted one place before a match occurs (*i.e.*, the input to the decision mechanism is $x_1x_2x_3x_4\,0\,1\,1$), a comparison looks like

$$\begin{array}{l} x_1\ x_2\ x_3\ x_4\ 0\ 1\ 1 \\ \qquad\uparrow\ 0\ 1\ 1\ 1\ 0. \end{array}$$

In this case, x_3 is in error. In general, all other information bits are correct.

One realization of the coder is shown in Fig. 1. One check digit is generated in each indicated path. All summers are modulo p adders or subtractors as shown. The gains a_i, being the proper multiplicative factors for the appropriate information digit, are time variant. The b_i's can be realized simply by switches that are open except at the time of the ith check bit output, $t = n - k - 1 + i$. This realization assumes the use of p-parity, *i.e.*,

$$\sum_{j=1}^{n} a_{ij}x_j + y_i = 0 \pmod{p}.$$

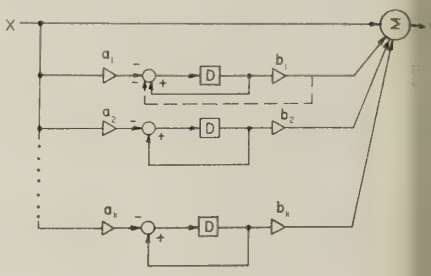


Fig. 1 Coder realization.

The decoder is identical in form to the coder and differs only in that a_i is replaced by $-a_i$.

For the $\mathbf{C}$ matrix of the example discussed above, the values of the multipliers are zero except

$$\begin{aligned} a_1 &= 1 \quad \text{at} \quad t = 0, 1, 3 \\ a_2 &= 1 \quad \text{at} \quad t = 0, 2, 3 \\ a_3 &= 1 \quad \text{at} \quad t = 1, 2, 3 \\ b_1 &= 1 \quad \text{at} \quad t = 4 \\ b_2 &= 1 \quad \text{at} \quad t = 5 \\ b_3 &= 1 \quad \text{at} \quad t = 6. \end{aligned}$$

The Hamming coder and decoder are not self-clearing (which is also true of any Huffman-type system). The system can be cleared without any loss in time in either of two ways. The first method is to place a normally closed switch in each feedback path. This switch would have the same control as the b_i (a normally open switch). The second method is to add a feedback path from the output of b_i to the summer (at the input of the delay) with a negative input as shown in Fig. 1 with a dashed line. This would be nonzero only at the time of the output of that particular check digit and would cancel the feedback from the solid path. Neither method affects the transmission matrix $\mathbf{H}$, since b_i is zero at all times after this clearing takes place. In normal operation, where the input is repetitive (*i.e.*, a new input occurs every n time units), this allows operation without interference among words.

The author wishes to acknowledge the many helpful discussions with Profs. B. Friedland and T. E. Stern.

Alan B. Marcovitz
Columbia University
New York, N. Y.

Contributors

Phillip Bello (S'52—A'55) was born in ...n, Mass., on October 22, 1929. He ...ived the B.S.E.E. degree from North...tern University, Boston, Mass., in 1953, ... the M.S. and D.Sc. degrees in 1955 ... 1959, respectively, from the Mass...usetts Institute of Technology, Cam...dge.

From 1955 to 1957, he was first research ...ociate and then assistant professor of ...mmunications at Northeastern Univer..., teaching courses in the area of linear ...em analysis and statistical communica... theory. He was also during this time ...olved in classified research in the area ...tatistical communication theory. From ...6 to 1958, he was employed by Dunn ...gineering Associates, Cambridge, where ... was engaged in analytical studies ...ociated with various aspects of statistical ...ar theory.

Since 1958, he has been employed at the ...plied Research Laboratory of Sylvania ...ctronic Systems, Waltham, Mass., work... primarily in the area of statistical ...mmunication theory.

Dr. Bello is a member of Tau Beta Pi, ...ma Xi, and Eta Kappa Nu.

❖

Charles R. Cahn (S'51—A'52—M'57) ...s born on October 7, 1929, in Syracuse, ... Y. He received the B.E.E. degree in ...9, the M.E.E. degree in 1951, and the ... D. degree in electrical engineering in ...5, all from Syracuse University, Syra...e, N. Y

From 1949 to 1956, he served first as ...tructor and later as assistant professor ...the Electrical Engineering Department ... Syracuse University, where he was ...gaged in research work in information ...ory, antenna theory, network theory, ...l systems engineering. From 1952 to ...3, he was employed in the System ...nning Department of the Niagara ...hawk Power Corporation, Buffalo, ...Y., where he was concerned with system ...nning and economic operation of a ...ge integrated power system. From 1956 ... 1959, he was on the staff of Ramo-...oldridge, a division of Thompson Ramo-...oldridge, Inc., Los Angeles, Calif., ...ere he worked on systems analysis and ...thesis, with emphasis on applications ...nformation theory in the field of digital ...mmunications. He has also investigated ...niques for electronic countermeasures ... methods of achieving reliable and high...ed transmission over fluctuating circuits.

... 1959, he joined Space Technology ...oratories, Inc., Los Angeles, as a ...mber of the Senior Staff in the Guidance ...earch Laboratory. He was concerned ...h the analysis and design of communica... systems for satellite applications, with ...phasis on methods for achieving trans...sion security.

...n 1960, he joined the Bissett-Berman ...poration, Los Amgeles, Calif., where ...is concerned with communications re...lated to intelligence and countermeasures problems.

Dr. Cahn is a member of Sigma Xi.

❖

Carl W. Helstrom, for a biography please see page 559 of the December, 1960, issue of these TRANSACTIONS.

❖

Richard M. Karp (S'58—M'60) was born in Boston, Mass., on January 3, 1935. He received the B.A. degree in mathematics in 1955, and the M.S. and Ph.D. degrees in applied mathematics in 1956 and 1959, respectively, all from Harvard University, Cambridge, Mass. He held a Harvard National Scholarship from 1951 to 1956. From 1957 to 1959 he was on the staff of the Harvard Computation Laboratory.

He has held summer positions with the General Electric Company and Lincoln Laboratory, Massachusetts Institute of Technology. Since January, 1959, he has been employed at the IBM Research Center, Yorktown Heights, N. Y., where he has done research in the theory of digital computer programming, information theory, and switching theory. He is also adjunct assistant professor of electrical engineering at New York University, New York, N. Y.

Dr. Karp is a member of Sigma Xi and RESA.

❖

William L. Kilmer (A'54) was born in Easton, Pa., on July 5, 1932. He received the B.S. and M.S. degrees in electrical engineering from the Pennsylvania State University, University Park, in 1954 and 1955, respectively, and the Ph.D. degree in electrical engineering from the University of Michigan, Ann Arbor, in 1958.

Nearly all of his work has been in the general area of digital computer design. Through the fall of 1958, he was successively employed at the Research and Development Center, Griffiss Air Force Base, Rome, N. Y.; Bell Telephone Laboratories Inc., Whippany, N. J., where he participated in the logical design of the Leprechaun Computer; The University of Michigan Engineering Research Institute, where he worked on high-speed computer components; and IBM Research, Yorktown Heights, N. Y. Since October, 1958, he has held an assistant professorship in electrical engineering at Montana State College, Bozeman, and a research appointment at the Montana State College Electronics Research Laboratory, where he is presently heading a computer research group to work in the areas of logical design procedures, digital coding, and abstract automata.

Dr. Kilmer is a member of the ACM, Sigma Xi, and several engineering honor societies.

Glenn M. Roe was born in Stanley, Wis., on December 17, 1916. He received the B.A. degree from St. Olaf College, Northfield, Minn., in 1938, and the M.A. and Ph.D. degrees in theoretical physics from the University of Minnesota, Minneapolis, in 1940 and 1947, respectively.

From 1941 until 1946 he was a physicist with the U. S. Navy Bureau of Ships, working on problems in hydrodynamics and underwater sound. He joined the Knolls Atomic Power Laboratory in 1947, and in 1954 transferred to the Electron Physics Research Department of the General Electric Research Laboratory, Schenectady, N. Y. His research interests have been in wave propagation, boundary value problems, neutron scattering and absorption, microwave theory, numerical methods, and automatic control.

Dr. Roe is a member of the American Physical Society, the Mathematical Association of America, the Society for Industrial and Applied Mathematics, AAAS, and Sigma Xi.

❖

George S. Sebestyen (M'55) was born in Budapest, Hungary, on November 27, 1931. He received the B.S. and M.S degrees in 1955, and the D.Sc. degree in 1959, from the Massachusetts Institute of Technology, Cambridge.

During his association with M.I.T., he served as research and teaching assistant. He joined the research department of Melpar, Inc., Boston, Mass. in 1955, where he was engaged in research and development in surveillance systems and in the application of statistical theory of communications to problems in secure radar and communication systems. Since 1957 he has engaged in research in the field of pattern recognition. Dr. Sebestyen is serving presently as project engineer at the Advanced Development Laboratory of Litton Systems, Inc., Waltham, Mass.

❖

Gerald M. White (S'56—M'58) was born in Detroit, Mich., on December 6, 1929. He received the B.S. degree in 1951 from the University of Michigan, Ann Arbor, and the M.S. and Ph.D. degrees in applied physics in 1953 and 1958, respectively, from Harvard University, Cambridge, Mass. During the academic year 1951–1952, he attended the Oak Ridge School of Reactor Technology, Oak Ridge, Tenn. He spent the academic year 1954–1955 at the Imperial College, London, England, under a Sheldon Traveling Fellowship.

In 1957 he joined the Information Studies Section of the General Electric Research Laboratory, Schenectady, N. Y., where he is presently doing research in noise theory and adaptive systems.

Dr. White is a member of Sigma Xi.

Abstracts

This Section of the issue is devoted to abstracts of material which may be of interest to PGIT members. Sources are Government, Industrial and University reports, and books and journals published outside of the United States. Readers familiar with material of this nature which is suitable for abstracting are requested to communicate the pertinent information to one of the Editors or Correspondents listed below.

A Statistic Associated with the Joint Distribution of N Successive Amplitudes—W. C. Hoffman (in English). (Dept. of Math., University of California, Los Angeles, Calif., Ph.D. dissertation; 1953.)

The joint distribution of n successive amplitudes refers to n values from the discrete-parameter stochastic process $R_j = \{X_j^2 + Y_j^2\}^{\frac{1}{2}}$ (J an integer), where $\{X_j, Y_j\}$ constitute a stationary Gaussian process. The R_j process has applications in electronics, though its sampling properties have not previously been investigated to any great extent.

The covariance matrix of the Gaussian process is defined, and certain of its properties explored. Two lemmas are proved concerning the inverse covariance matrix and the eigenvalues of the covariance matrix. A useful property of a parameter of the R-distribution is determined and the covariance function of the R_j process found. The bivariate, trivariate, and joint n-dimensional probability density functions of the R_j process are then derived, using the lemma on the inverse covariance matrix and some results from the theory of Bessel functions.

A statistic q, consisting of the mean value of the sum of n successive values of the square of the R_j process, is defined next, and shown to be an unbiased and consistent estimator of a parameter of the R-distribution. The characteristic function of q is found by a modification of the diagonal elements of the inverse covariance matrix of the Gaussian process and an application of a basic property of probability density functions. After some algebraic reductions, the characteristic function can be put in a form amenable to the Fourier inversion formula. Evaluation of the latter by the calculus of residues yields the small sample distribution of q in terms of the eigenvalues of the covariance matrix of the Gaussian process.

It is then shown that the statistic q, under a hypothesis that amounts essentially to the existence of a spectral density function for the Gaussian process, is asymptotically normally distributed. In the case of simple Markov dependence, the hypothesis of the theorem is automatically satisfied.

A test is prescribed for independence versus simple Markov dependence for the R_j process, assuming the availability of independent realizations. This test is equivalent, in the simple Markov case at least, to the uniformly most powerful one-sided test of the variance in random sampling from a normal distribution. The power function of the test is determined and depicted graphically for several sample sizes at the 95 per cent significance level.

The Markov case of the R_j process is studied in some detail. Among other results, it is shown that the process is ergodic. Formulas are given for maximum likelihood estimates of the parameters of the transition density functions. Using Wald's and Kazami's results on the asymptotic properties of maximum likelihood estimates in the case of dependent random variables, the asymptotic efficiency of the statistic q is determined in the case of simple Markov dependence.

Summary of Maximum Theoretical Accuracy of Radar Measurements—R. Manasse (in English). (Mitre Corp., Lexington, Mass., Tech. Series Rept. No. 2.)

This paper summarizes some general formulas for maximum theoretical accuracy of radar measurements on a target in the presence of additive white Gaussian noise. The formulas are specialized to some particular cases of interest.

Parameter Estimation Theory and Some Applications of the Theory to Radar Measurements—R. Manasse (in English). (Mitre Corp., Lexington, Mass., Tech. Series Rept. No. 3; no date.)

The general theory of parameter estimation is developed using the inverse probability approach. Where the measurements are perturbed by additive Gaussian noise, and when the received information is sufficient to determine the parameters of interest rather accurately, it is shown that an optimum method of processing redundant data based on the maximum likelihood approach reduces approximately to the solution of **k** nonlinear equations in the **k** unknown parameters. An expression is derived for the resulting error moment matrix of the parameters. It is shown that this same moment matrix for a minimum variance estimate is obtained by using results derived by Cramér. The theory is illustrated by applying it to several radar measurement problems of interest.

Theory and Application of the Separable Class of Random Processes—A. H. Nuttall (in English). Res. Lab. of Electronics, Mass. Inst. Tech., Cambridge, Mass., Tech. Rept. No. 343; May 26, 1958.)

The separable class of random processes is defined as that class of random processes for which the g-function,

$$g(x_2, \tau) = \int_{-\infty}^{\infty} (x_1 - \mu) p(x_1, x_2; \tau)\, dx_1$$

separates into the product of two functions, one a function only of x_2, the other a function only of τ. The second-order probability density function of the process is $p(x_1, x_2; \tau)$ and has μ as its mean. Various methods of determining whether a random process is separable are developed, and basic properties of the separable class are derived.

It is proved that the separability of a random process that is

sed through a nonlinear no-memory device is a necessary and icient condition for the input-output crosscovariance function e proportional to the input autocovariance function, whatever linear device is used. The uses of this invariance property are ted out.

a nonlinear no-memory device is replaced by a linear memory- able network, so as to minimize the mean-square difference ween the two outputs for the same separable input process, lysis shows that the optimum linear network has no memory. ple relations among correlation functions for these circuits are derived.

ome results on Markov processes and best estimate procedure derived, important examples of separable processes are given, possible generalizations of separability are stated.

liometer Techniques in Radar—R. Price (in English). (Lincoln ., Mass. Inst. Tech., Lexington, Mass., Group Rept. 34G-0003; e 10, 1960.)

gated radiometer is described that is appropriate to the detection adar echoes from a scintillating target whose dynamical behavior ell known, when the echoes are submerged in a strong back- und of white noise. A possibly novel feature of the gated radio- er is that simply by adjusting its predetection filter it can be le to perform radar sweep integration that is purely predetection, ely postdetection, or a mixture of both. The close relationship he gated radiometer to sweep integration and conventional radio- try is discussed.

The output signal-to-noise ratio of the gated radiometer is de- d and special cases of interest are examined. Range-frequency biguity behavior is studied, and the advantages of coding the smission are mentioned. A proof is given that under certain ditions, which are probably quite realistic, the gated radiometer n optimum detector.

Finally, recent applications of gated radiometer techniques are cribed in part. Radar echoes from Venus, from the Sun, and from electrons in the ionosphere have been detected by such means, , in addition, new information has been gained about the moon.

timum Codes Study—R. Turyn (in English). (Applied Res. Lab., vania Electronic Systems, Waltham, Mass., Final Rept.; January 1960.)

This report is concerned with finding sequences whose terms are or -1, or codes, and whose autocorrelation function is relatively ll (except at zero shift). The main interest is in certain optimal codes (autocorrelation function not greater than 1 in absolute value except at zero shift). The main results are that no such optimum codes exist if of odd length exceeding 13. For even length, certain periodic codes are considered and some important results are given. As a consequence, no optimum codes of even length less than 144 but greater than 4 exist. It thus seems likely no optimum codes other than those known exist. In addition, certain near-optimum codes are discussed.

Phase-Shift Keying in Fading Channels—H. B. Voelcker (in English). (*Proc. IEE*, vol. 107, pt. B, pp. 31–38; January, 1960.)

Phase-shift keying (psk) is discussed as a modulation technique for transmitting digital data over radio circuits subject to fading. The modest bandwidth requirements of psk modulation suggest that it can not only alleviate spectrum crowding, but can also transmit traffic with fewer errors. The theoretical results presented here indicate, however, that the random phase perturbations inherent in fading radio signals cause unavoidable degradation in the performance of psk systems. Experimental data which partially support the theoretical results are cited, and comments relevant to the improvement of psk systems are included.

A Theory of Signals—R. E. Wernikoff (in English). (Res. Lab. of Electronics, Mass. Inst. Tech., Cambridge, Mass., Tech. Rept. No. 331; January 31, 1958.)

An experiment is presented in which an attempt is made to arrive at a mathematical description of physical signals that embodies, more realistically than the usual functional representation, our limitations in performing measurements. The object is to achieve a closer relation between the structure of the mathematical description and the finite-resolution properties of the detector that characterize any real measurement process.

An algebra of signals is obtained, appropriate to the model in which essentially frequency-limited signals interact with linear, time-invariant systems, and observations are made by means of a linear, finite-resolution oscilloscope. The properties of this algebra are studied, and a metric that indicates which operators give physically indistinguishable outputs is defined. The algebra is used to study problems in uniform and nonuniform sampling, the discrimination of two events from one in noisy, radar-like systems, and the conditions under which a signal is indistinguishable from its short-time average. A general procedure for linear, least-peak-error prediction is obtained. In the limit as the detector resolution becomes perfect, the present model is shown to tend smoothly to the usual functional model.

ook Reviews

tistical Theory of Signal Detection—Carl W. Helstrom, Ph.D. rgamon Press, Inc., New York, N. Y.; 1960. 334 pages. $9.50.)

This book is intended as an interdisciplinary text for the mathe- tician and radar or communications engineer, to introduce statis- l decision theory and to apply it to problems in signal detection. fulfilling his intention "to convey to each some feeling for the ject, so that each can exploit it for his own purposes," the author produced a very readable volume. Of special merit are the ny discussions of the usefulness and meaning of assumptions and hods of solution, and the summary on page 330. The detail ded for completeness and rigor has naturally been omitted; these ssions are noted in the text and references are adequately given. The applications treated are detection of radar and digital com- nication signals in added Gaussian noise. Both white noise and se with known autocorrelation are considered, and the integral ations and solutions necessary to obtain orthonormal samples treated.

The first two chapters fill in necessary background on signals, ar filters, and noise. Chapter three introduces decision theory its methods, covering all of the usual approaches including uential analysis, though the remainder of the work concentrates fixed observation time decisions based on a Neyman-Pearson erion. Subsequent chapters deal with the completely known signal, and then progressively with ensembles with unknown amplitude, time of arrival, and pulse-to-pulse fluctuations. A good development in chapter five shows how minimax approaches and small signal approximations can lead to unsatisfactory design by concentrating on weak and useless signals. The theory of estimation is introduced and discussed, and applied to the detection problems for unknown parameters where the assumption of least-favorable values is impractical. In addition there are chapters treating multiple signals and detection in clutter, and the detection of stochastic signals.

In keeping with the intention to convey feeling for the subject, a discussion on the physical implication of the integral equation-orthonormalization technique as "whitening," and the singular cases arising from it would have been a desirable addition.

This book should be useful not only as an introduction, but also to compliment, on one hand, the more practical literature, and on the other hand, the more rigorous. It is recommended to both the graduate student and to the worker who is trying to understand the use of statistical theory in receiver design.

Prof. T. G. Birdsall
Cooley Electronics Labs.
University of Michigan Res. Inst.
Ann Arbor, Mich.

A STATEMENT OF EDITORIAL POLICY

The IRE TRANSACTIONS ON INFORMATION THEORY is a quarterly journal devoted to the publication of papers on the transmission, processing, and utilization of information. The exact subject matter of acceptable papers is intentionally, by editorial policy, not sharply delimited. Rather, it is hoped that as the focus of research activity changes, a flexible policy will permit the TRANSACTIONS to follow suit and that it will continue to serve its readers with timely articles on the fundamental nature of the communication process. Topics of current appropriateness include the coding and decoding of digital and analog communication transmissions, studies of random interferences and of information bearing signals, analyses and design of communication and detection systems, pattern recognition, learning, automata, and other forms of information-processing systems.

Papers can be of two kinds, tutorial or research, and should be so indicated. The former must be well-written expositions summarizing the state of a field in which research is still in progress, or else unifying results scattered in the literature. Research papers must be original contributions not published elsewhere. They must present new methods, concepts, or ideas, or extend old ones to new areas of applicability; or, they must present new data, findings or inventions, or solve new problems of more than casual interest. They will not be accepted if, in the view of the reviewers and editors, they constitute a straightforward and easy application of existing theory to a special case of limited interest. It is not necessary that the length of each research paper be great; on the contrary, the submission of short but formal research notes is to be encouraged.

In addition to papers, readers are invited to submit notes to the Correspondence section. These may include such things as early summaries of important work to be published later at greater length, or remarks on material that has already appeared. Contributions in the form of "problem statements" are also sought for the Correspondence section. This category includes problems to which the author knows no solution but suspects that another reader might, conjectures for which a proof or disproof is desired, and so forth.

INFORMATION FOR AUTHORS

Authors are requested to submit editorial correspondence or technical manuscripts to the Editor for possible publication in the PGIT TRANSACTIONS. Papers submitted should include a statement as to whether the material has been copyrighted, previously published, or submitted for publication elsewhere.

To expedite reviewing procedures, it is requested that authors submit the original and two legible copies of all written and illustrative material. The manuscript should be double-spaced, and the illustrations drawn in India ink or drawing paper or drafting cloth. Each paper should include a carefully written abstract of not more than 200 words. Papers should be prepared for publication in a matter similar to those intended for the PROCEEDINGS OF THE IRE. Further instructions may be obtained from the Editor. The original copy and drawings of material not accepted for publication will be returned.

All technical manuscripts and editorial correspondence should be addressed to Arthur Kohlenberg, Melpar, Inc., 11 Galen Street, Watertown 72, Mass.

Local Chapter activities and announcements, as well as other nontechnical news items, should be addressed to the PGIT Newsletter, c/o Prof. N. M. Abramson, Electrical Engineering Department, Stanford University, Stanford, Calif.

INSTITUTIONAL LISTINGS

The IRE Professional Group on Information Theory is grateful for the assistance given by the firms listed below and invites application for Institutional Listing from other firms interested in the field of Information Theory.

IBM RESEARCH, INTERNATIONAL BUSINESS MACHINES CORP., Yorktown Heights, N. Y.
Error Correcting & Detecting Codes, Theory of Assemblies & Automata, Information Networks, Reliability

REPUBLIC AVIATION CORP., Farmingdale, N. Y.
Aircraft, Missiles, Drones, Electronic Analyzers; U. S. Distr. of Alouette Turbine-Powered Helicopter

The charge for an Institutional Listing is $50 per issue or $150 for four consecutive issues. Applications for Institutional Listings and checks (made payable to the Institute of Radio Engineers) should be sent to L. G. Cumming, Institute of Radio Engineers, 1 East 79 St., New York 21, N. Y.